Konputazio Zientziak eta Adimen Artifizialaren Saila
Departamento de Ciencias de la Computación e Inteligencia Artificial

Data analysis advances in marine science for fisheries management: Supervised classification applications

by

Jose Antonio Fernandes Salvador

Supervised by Jose Antonio Lozano, Iñaki Inza and Xabier Irigoien

Dissertation submitted to the Department of Computer Science and Artificial Intelligence of the University of the Basque Country as partial fulfilment of the requirements for the PhD degree in Computer Science

Donostia - San Sebastián, June 29, 2011

Summary

The impact of how fisheries are managed is of great importance on biological, economic, social and political levels. However, there is still a high uncertainty about the relationships between climate, fish and management decisions. Many activities are performed in marine science in order to reduce this uncertainty. This dissertation provides methodological contributions to several of the activities necessary for fisheries management, with three main contributions and several related minor contributions.

Firstly, this dissertation deals with the challenge faced by experts when using supervised classification in order to process the increasing number of samples (e.g. scanned zooplankton). This challenge consists of a trade-off between the performance and the number of zooplankton taxa to classify; i.e. usually the higher the number of classes or taxa, the lower the performance. The contribution in this domain is a wrapper method where the expert can evaluate the training set in terms of this trade-off between performance and the number of taxa to classify. In relation to this topic, other minor contributions have been accomplished in the classification of phytoplankton, otoliths and habitats.

Secondly, the problem of robust forecasting in domains of scarce data, such as fish recruitment forecasting, is dealt with. A methodological pipeline of machine learning state-of-the-art methods is proposed and its proper application is verified. The proposed methodology allows building a probabilistic model where three levels of anchovy recruitment (low, medium and high) can be predicted based on a small set of factors. In addition, the methodology allows identifying relevant boundaries of recruitment levels given environmental factors, as well as a small set of factors that tend to have a low correlation between them and be highly correlated with the recruitment. Finally, this work has triggered several collaborations where the modelling is adapted and combined in an extended framework to deal with the practical necessities of fisheries management.

Finally, the new machine learning paradigm of multi-dimensional classifiers is applied to simultaneous multi-species recruitment forecasting in the context of an ecosystem-based approach to fisheries management. Multi-dimensional classifiers aim to perform the simultaneous forecasting of several variables, which suits the need of simultaneously forecasting several fish species that share ecosystems. The study proposes a set of pre-processing methods adapted to the multi-dimensional approach by combining them with multi-dimensional classifiers. The proposed multi-dimensional pre-processing methods are tested on both synthetic and real domains. In addition, an extensive comparison among proposed multi-dimensional pre-processing methods is provided, as well as an analysis of under which circumstances each method can be superior to the rest.

Resumen

La gestión de pesquerías tiene un gran impacto a muchos niveles: biológico, económico, social y político. Sin embargo, aun hay mucha incertidumbre sobre las relaciones entre el clima, los peces y las decisiones de gestión. Muchas actividades son realizadas en investigación marina para reducir esta incertidumbre. En esta tesis doctoral varias contribuciones metodologícas son presentadas en torno a estas actividades necesarias para la gestión de pesquerías, con tres contribuciones principales y varias aportaciones adicionales.

En primer lugar, esta tesis se enfrenta al desafío que se presenta a muchos expertos cuando quieren aplicar clasificación supervisada para procesar el cada vez mayor numero de muestras biológicas. Este desafío consiste en encontrar un equilibrio entre el número de classes a classificar y el rendimiento de la clasificación. Normalmente, a un mayor número de clases el rendimiento es menor. La contribución en este dominio consiste en un método 'wrapper' donde el experto puede evaluar el conjunto de entrenamiento en base a este equilibrio entre rendimiento de la clasificación y el número de clases a distinguir. Otras contribuciones, relacionadas con este tema, han sido realizadas en las áreas de clasificación de fitoplancton, otolitos y hábitats.

En segundo lugar, se propone una metodología para afrontar el problema de predecir robustamente en un dominio con pocos datos como es la predicción del reclutamiento de peces. La metodología consiste en la propuesta de un grupo secuencial de reconocidos métodos provenientes del aprendizaje automático y como aplicarlos correctamente. Esta metodologá permite aprender un modelo probabilístico donde tres niveles de reclutamiento (bajo, medio, alto) se pueden predecir. Además, la metodología permite restringir el número de factores a usar en la predicción que tienden a estar poco correlacionados entre ellos y muy correlacionados con el reclutamiento simultaneamente Adicionalmete, permite identificar las fronteras entre los distintos niveles de reclutamiento dado ese restringido número de factores. Finalmente, este trabajo a dado lugar a varias colaboraciones donde el modelo ha sido adaptado y combinado con otros según de cara a una mejor gestión de pesquerías.

En tercer lugar, el nuevo paradigma de clasificadores multi-dimensionales es aplicado a la predicción simultanea de multiples especies de peces en un contexto de una aproximación ecosistémica a la gestión de pesquerías. Los clasificadores multi-dimensionales tienen este objetivo de hacer predicción simultanea de varias variables objetivo, lo que encaja con el objetivo de clasificar simultaneamente varias especies de peces que comparten ecosistema. En esta tesis la adaptación de un grupo de métodos para pre-proceso de los datos es propuesto considerando el objetivo multi-dimenional y para combinarlos con clasificadores multi-dimensionales. Los metodos de pre-proceso multi-dimensionales propuestos son verificados con datos reales y con datos artificiales. Además, una extensa comparación de los métodos es realizada, asi como un análisis de en que circunstancias un método en particular muestra un comportamiento superior al resto.

Laburpena

Arrantza kudeaketaren eragina oso garrantzitsua da arlo biologiko, ekonomiko, sozial nahiz politikoan. Hala ere, klima, arrantza eta kudeaketa erabakien arteko erlazioenganako zalantza handiak ageri dira. Aipatu zalantzak argitzeko asmoz hainbat aktibitate burutzen dira itsas zientziaren baliabideetan. Honako tesi honek arrantza kudeaketarako beharrezkoak diren aktibitateentzat hiru ekarpen metodologiko nagusi aurkezten ditu, beste hainbat elementurekin batera.

Lehenik eta behin, tesi honek lagin kopuru handiak (adibidez, eskanetautako zooplankton laginak) prozesatzeko erabiltzen den sailkapen ikuskatua aztertzen du. Erronka hau, prozesuaren adierazgarritasun eta sailkatzeko zooplankton taxoi kopuruaren arteko orekatzean datza; hau da, normalean gero eta taxoi edo klase kopuru handiagoa izan, are eta txikiagoa izango da emaitzaren adierazgarritasuna. Gai honen ekarpena hau konpontzen saiatzeko metodo berri batean datza, non adituak entrenamendu edo Training Set-a ebaluatu ahal izango duen aurrerago aipatu oreka horren baitan. Gai honekin erlazionaturik, beste hainbat ekarpen ere egin dira, hala nola, fitoplanktonaren sailkapenean edota otolito eta habitatetan.

Bigarrenez, arrainen erreklutamenduaren iragarpena bezalako datu urriko domeinuetan iragarpen sendoak egiteko arazoa aztertzen da. Goi mailako makina-ikasketa arloan ekarpen metodologiko bat egiten da, honen aplikazioaren egokitzapena egiaztatzen delarik. Proposatu metodologia honek modelo probabilistiko bat eraikitzea baimentzen du, non faktore sorta txiki batean oinarriturik antxoaren hiru erreklutamendu maila (baxua, ertaina, altua) iragarriak izan daitezken. Honetaz gain, metodologiak erreklutamendu mailen muga aipagarriak nahiz emandako ingurumen faktoreak identifikatzen laguntzen du, haien artean korrelazio-maila baxua izanik ere, erreklutamenduarekin estuki korrelazionaturiko faktore sorta txiki batekin batera. Azkenik, lan honek hainbat elkarlan ahalbidetu ditu, non modeloa marko zabal batean moldatua eta konbinatua izan den arrantza kudeaketarako beharrak asetzeko asmoz.

Hirugarrenez, sailkatzaile multi-dimentsionalek osatzen duten makina-ikasketa arloko paradigma berria erreklutamendu multi-espezifikoen iragarpenetan aplikatua izan da, ekosisteman oinarrituriko arrantza kudeaketaren kontextuan. Sailkatzaile multi-dimentsionalek hainbat aldagaien aldi bereko iragarpena dute helburu, eta honek ekosistema berean bizi diren hainbat arrain espezie aldi berean aztertzea eskatzen du. Ikerketa honek hurbilketa multi-dimentsionalerako aurre-prozesatze metodo sorta bat proposatzen du, metodo hauek sailkatzaile multi-dimentsionalekin batera konbinatuz. Proposaturiko aurre-prozesatze metodo multi-dimentsionalak domeinu sintetiko zein errealetan testatuak izan dira. Honetaz gain, proposaturiko aurre-prozesatze metodo multi-dimentsionalen arteko konparaketa zabala aurkezten da, metodo bakoitza besteegandik zein egoeratan gailenduko den aztertzeko analisi batekin batera.

Résumé

La gestion des pêches a un grand impact biologique, économique, social et politique. Cependant, même beaucoup d'incertitudes existent sur les relations entre le climat, des poissons et les décisions de gestion. De nombreuses activités sont menées en recherche marine pour réduire cette incertitude. Dans cette thése de doctorat plusieurs contributions méthodologiques sont présentées autour de ces activités nécessaires pour la gestion des pêcheries, avec trois contributions principaux et Plusieurs contributions supplémentaires.

En premier lieu, cette thse est confronté au défi qui se présente a des nombreux experts lorsqu'ils veulent appliquer le classement supervise pour traiter le plus en plus grand nombre d'échantillons biologiques. Ce défi consiste à trouver un équilibre entre le nombre de classes à classifier et le rendement de la classification. Normalement, à un plus grand nombre de classes le rendement est plus petit. La contribution dans ce domaine consiste a une méthode 'wrappe' où l'expert peut évaluer l'ensemble de formations sur la base de cet équilibre entre le rendement de la classification et le nombre des classes distinguer. Autres contributions, liées ce point, ont été menées dans les domaines de classement de phytoplancton, otolithes et habitats.

Deuxiémement, il est proposé une méthodologie pour affronter le probléme de faire une prédiction robuste dans un domaine avec quelques données disponibles comme c'est la prédiction du recrutement de poissons. La Methodologie consiste la proposition d'un groupe séquentiel de méthodes reconnues provenantes de l'apprentissage automatique et comment les appliquer correctement. Un modéle probabilistique permet d'apprendre cette méthodologie o trois niveaux de recrutement peuvent être prédits. La méthodologie permet de restreindre le nombre de facteurs à utiliser dans la prédiction qui tendent à être en rapport entre ceux-ci et trés mis en rapport avec le recrutement simultanément. Permet d'identifier les frontiéres entre des niveaux distincts de recrutement qui ont donné ce restreint par un nombre de facteurs. Ce travail à un lieu donné pour quelque collaboration où le modèle a été adapté et combiné par les autres selon vis-à-vis une meilleure gestion de pêcheries.

Finalement, le nouveau paradigme de classificateurs multi-dimensionnels est appliqué à la prédiction simultanée d'espéces multiples de poissons dans un contexte d'une approche écosystématique la gestion de pcheries. Les classificateurs multi-dimensionnels ont cet objectif de faire une prédiction simultanée de différentes variables, ce qui s'embote avec l'objectif de classer simultanément quelques espéces de poissons qui partagent un écosystème. Dans cette thse l'adaptation d'un groupe de méthodes pour pré-processus des données est proposée en considrant l'objectif multi-dimensionnel d'où on peut les combiner avec des classificateurs multi-dimensionnels. Les méthodes multi-dimensionnelles de pré-processus proposées sont vérifiées par des données réelles et autres artificielles. Une comparaison étendue des mthodes est réalisée, ainsi que une analyse a toujours des circonstances il y a une méthode qui montre en particulier un comportement supérieur au reste.

Acknowledgements

I am grateful to my supervisors Jose Antonio Lozano, Iñaki Inza and Xabier Irigoien for their patience and advice as well as for their effort not only when writing this PhD dissertation, but also in making me a good researcher as well.

I am also grateful to Fundación Centros Tecnologicos Iñaki Goenaga for the opportunity they have given me, and to AZTI-Tecnalia for creating an environment which allowed the writing of this thesis. An environment where people do not know the meaning of the words 'no', 'I'm busy' or 'I can not'. Thanks to the many people whose previous work has contributed to this thesis and many that just supported or helped me or read and commented on this dissertation. Among them are Maria Santos, Leire Ibaibarriaga, Angel Borja, Andres Uriarte, Javier Franco, Guillermo Boyra, Eneko Bachiller, Ainhoa Lezama, Nerea Goikoetxea, Aitor ascoreca, Sonia Sanchez, Maitane Grande, Aitor Albaina, Maria Korta, Ainhoa Lezama and others.

I am also grateful to all the ISG group for their support and help. In particular, to Juan Diego Rodríguez and Aritz Pérez due to their dedication, time and patience with my questions. I also have to thank Borja Calvo for his help with some 'latex' issues during the writing of this dissertation.

I would also like to thank several foreign researchers, who I have visited, for their teaching and kindness: Philippe Grosjean and Kevin Denis (Mons-Hainaut University) in Belgium; Sakari Kuikka, Laura Uusitalo, Kirsi Hoviniemi and Teppo Juntunen in Finland. In particular, the research in plankton classification, and some hints of hot topics in that area, are due to the dedication and discussion provided by Philippe Grosjean. I have also had some relationship with other research centers and their staff: Miguel Alcaraz, Enric Saiz and Patricia Jiménez of the CSIC (Spain); Antonio Bode of the IEO (Spain) and; Vivi Fleming, Heikki Pitkänen and Päivi Korpinen of the SYKE (Finland). Thank very much for your time and dedication.

The present dissertation and other researches would not be possible without the work of many people who usually remain unknown: laboratory analysts as well as the crews of fisheries and oceanographic vessels. I have nice memories of my work in the vessels 'Investigador' and 'Garcia del Cid'. I have also fond memories of my time spent working with many analysts such as Naiara Serrano, Maite cuesta or Irene Gómez.

The writing of this dissertation has been possible thanks to the 'latex environment'. I have to thank the people that first introduced me to latex 10 years ago in the E-ghost group of the University of Deusto (Spain): Borja Sotomayor and Pablo Pérez. I am also grateful to many people who have contributed with samples such as Brian Amberg, Rasmus Pank, Till Tantau or Kjell Magne.

Finally, this dissertation would not have been possible without the financial support offered by several research projects, namely: 1) Basque Country Government funded projects ECOANCHOA, K-EGOKITZEN, Etortek and

Saiotek; 2) The Spanish projects PERFIL and Consolider Ingenio project of the Spanish Ministry of Science and Innovation, as well as COMBIOMED network in computational biomedicine of Carlos III Health Institute and, 3) The European projects FACTS, UNCOVER, and MEECE.

Above all, I would like to give my thanks to my family for their understanding and support throughout these years, in particular, during my breaks from the real world when I often needed to focus on my research work. I am also grateful to all my friends, who often lost contact with me but when we meet they are still the same friends: to all my old and new friends, to my dancing friends (and students), to my friends from the theatre group 'Cuerpo Escena' and, in particular, to our director Paulo Geiger who taught me to move columns and walls. I would like to give a special mention to Penkat Silat as well as the people that introduced me to it. I think that Silat training has provided me with the physical and mental strength needed to deal with the big task of writing a thesis in the hard times we live in.

I would also like to thank my ex-girl friend. Although we want different things from life, we went through hard times together and it would have been too difficult to go though them if we had not been together. I hope she can find what she needs in life. On the other hand, I feel very lucky that I found 'minum ilona'.

There are many other people who I just can not recall now that should be mentioned here. To all of them: thank you and forgive me for my bad memory.

Contents

List of Figures

List of Tables

Part I

Introduction

1

Introduction

The difficulties for direct observation of biological interactions and mechanisms have pushed marine science towards the development of large bodies of measurements from where the underlying mechanisms can be deduced (Duarte, 2007). In fact, major international programs (IGBP, JGOFS, GLOBEC, ICES and others), engaging thousands of marine scientists throughout the world over the past decade, have delivered a massive amount of information on the biogeochemical foundations, functioning and structure of marine food webs. Parallel technological developments, ranging from satellite imagery to autonomous underwater vehicles, have increased by orders of magnitude the resolution and amount of data available on relevant properties of the ocean ecosystem.

The resulting data represent a key resource to explore patterns in the structure and functioning of the ocean ecosystem that is yet to be fully utilised. Furthermore, the management of marine exploited resources (e.g. fisheries management) is also based on the collection of large amounts of data and the time series of catches and age structure by means of biological samples collected during oceanographic surveys (Fig. 1.1).

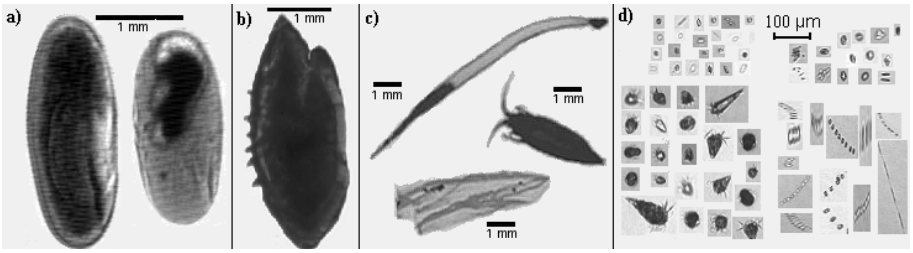


Fig. 1.1. Examples of biological samples collected during oceanographic surveys: a) anchovy eggs digitalized using Flowcam system; b) otolith, a bone in the head of a fish whose rings provide knowledge of their age; C) big zooplankton; and d) big phytoplankton

Surveys at sea are extremely expensive due to the cost of vessels and human resources. Therefore, the effort of samples and data collection during oceanographic surveys is maximized. However, those samples are often only partially analysed due to the lack of resources for the full analysis of all the samples collected. This results in a massive increase in the amount of data available that is costly to process and analyse, making knowledge extraction a challenge that enhances the use of data mining and semi-automatic techniques.

Bioman survey (Motos, 1996), performed by the marine research institution AZTI-Tecnalia (*http://www.azti.es*) every year in May, is an example of such surveys. Its main objective is to evaluate anchovy biomass using the egg production method. In this survey the main activity is to count anchovy eggs. However, many different types of biological samples (anchovy tissues, phytoplankton, mesozooplankton), physical (temperature, salinity, vertical structure) and environmental information are collected, using different kinds of instruments (e.g. Fig. 1.2).

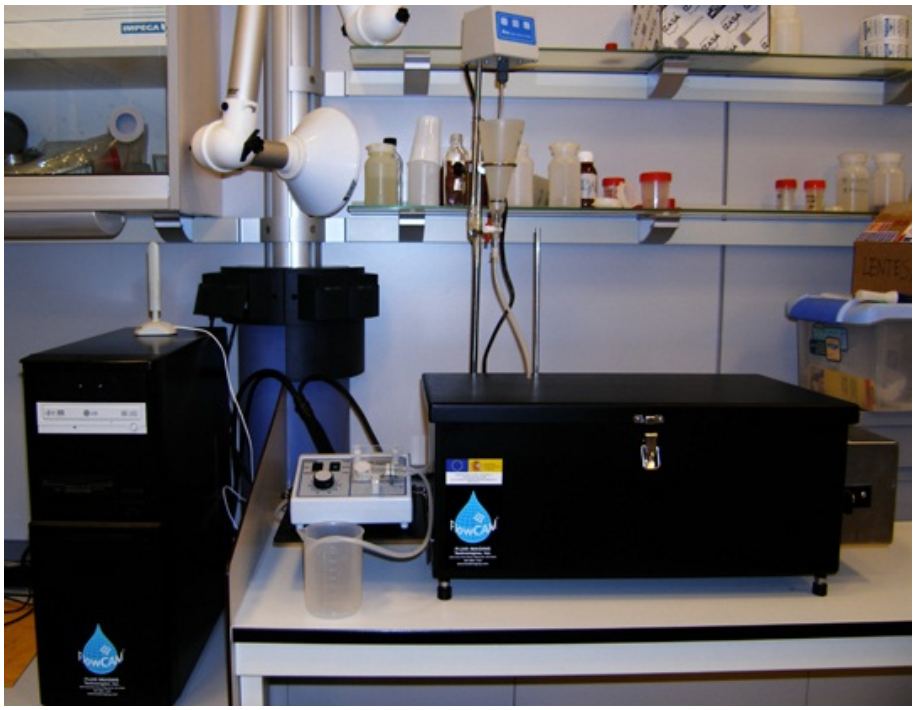


Fig. 1.2. Flowcam, instrument used for phytoplankton digitalization during oceanographic surveys or in laboratories. It is composed of a water plumb, a plastic pipe where the sample goes through and where a digital camera takes several photos per second of the particles in the water sample. Those particles often have sizes inferior to one millimeter. They are later counted and classified using *supervised classification* methods.

The aim of collecting these samples and environmental data is to improve the knowledge of oceanic ecosystems. One of the main uses of that knowledge is fisheries management (Motos and Wilson, 2006). Fisheries management deals with the control of the exploitation of fish species of commercial interest. One important aspect for fisheries management is the amount of new fish that enter the fishery each year (recruitment).

Actually in short lived species such as anchovy, recruitment determines the abundance of the population (Ibaibarriaga et al., 2008). Recruitment is dependent on the size of the parental population (number of eggs produced) (Ibaibarriaga et al., 2010), but mainly on the survival of those eggs depending on the environmental conditions (temperature, transport, predation, food) (Ricker, 1954; Cushing, 1982; De Oliveira et al., 2005). Therefore, the ability of processing massive amounts of samples, such as zooplankton (Zarauz et al., 2008; Irigoien et al., 2009), fitoplankton (Zarauz et al., 2009; Denis et al., 2009), eggs and larvaes (Motos, 1996; Ibaibarriaga et al., 2010), otolith (Ascoreca et al., 2008) and others, is crucial to understand some of those dependencies and relationships.

In addition, by definition, recruitment can only be measured once the individuals have entered the fishery. Being able to forecast the strength of the oncoming recruitment based on the size of the parental stock and the environmental variables would allow for a much better management of the fisheries, in particular for short lived species (Myers et al., 1995). Fisheries management can benefit of machine learning techniques that are especialized in dealing with *uncertainty* (Fernandes et al., 2010c).

In this dissertation several data domains related with the activities needed for fisheries management are addressed. These activities can be grouped in two categories from a data analysis point of view. Firstly, the activities related with the processing of samples deal with massive amounts of data. Secondly, environmental relationships with fish recruitment are difficult to identify due to the sparse data available (yearly averages). In addition, the available factors have a high level of noise since they are often collected by indirect measurements or they are just estimations. Therefore, recruitment forecast is a difficult problem since it shows sparse data and high *uncertainty*.

The objectives of this dissertation are to apply *supervised classification* techniques to these two groups of data analysis activities that are involved in the process of reducing *uncertainty* for improvement of fisheries management. Its aim is to improve semi-automatic processing of samples and the learning of robust recruitment forecasting models. The objective has not been just to apply available machine learning methods without adapting them to the domain. Therefore, these problems have been analysed to detect weaknesses in previous approaches that can be solved with methodological contributions.

1.1 Contributions of the dissertation

The contributions of this dissertation are both in the methodologies and their applications. The objective has been to develop novel classification methodologies specially designed for a set of marine science domains. These contributions are presented as three groups of work (Fernandes et al., 2009c, 2010c,b). A brief explanation of each one is given in the following paragraphs. Section 1.2 includes the full thesis overview, pointing to the particular chapters and sections where each item is presented and discussed.

A. Optimizing the number of classes in zooplankton classification

Zooplankton biomass and abundance estimation, based on surveys or time-series, is carried out routinely in marine research facilities (Irigoien et al., 2002, 2004). The analysis of those samples is costly in time (Boyra and Arregi, 2005). Therefore, automated or semi-automated image analysis processes, combined with machine learning techniques for the identification of plankton, have been proposed to assist in sample analysis (Culverhouse et al., 1996, 2003; Grosjean et al., 2004).

A difficulty in automated plankton recognition and classification systems is the selection of the number of classes (Hu and Davis, 2006; Fernandes et al., 2009c). This selection can be defined as a balance between the number of classes identified (zooplankton taxa) and performance (accuracy; correctly classified individuals) (Fernandes et al., 2009c).

A method is proposed to evaluate the impact of the number of selected classes, in terms of classification performance (Fernandes et al., 2009c). On the basis of a dataset of expert labelled zooplankton images, a machine learning method suggests groupings of classes that improve the performance of the automated classification. The end-user can accept or reject these mergers of classes depending on their ecological value and the objectives of the research.

This method allows both objectives to be balanced: a) maximization of the number of classes; and b) performance, guided by the end-user. This study allowed processing thousands of images that led to another important biological contribution (Irigoien et al., 2009).

B. Fish recruitment forecasting using robust supervised classification methods

Many studies have been undertaken on the environmental and climatic factors that influence recruitment of different fish species (Ricker, 1954; Cushing, 1982; Myers et al., 1995). The interactions are complex and often non-linear; such that, frequently, the different factors are difficult to disentangle (Myers et al., 1995; Schirripa and Colbert, 2006; Planque and Buffaz, 2008).

The main difficulty in this domain is to learn a reliable model due to the sparse and noisy nature of the available data (Schirripa and Colbert, 2006). In management context, the accuracy of the forecast has important consequences and, to be useful, the manager needs to know the risk that is being taken. As a result, the objective of this study has been to build a robust classifier for fish recruitment forecast associated to its risk or *uncertainty*.

The proposed methodology consists of a pipeline of methods (Fernandes et al., 2010c): a novel semi-automated recruitment discretization method (discretization of the class variable); factors supervised discretization; multivariate and non-redundant feature selection; and a final *naive Bayes* classifier. *Bayesian network classifiers* such as *naive Bayes*, have the advantage that they not only provide forecasts, but also the estimated probability of each possible outcome. The robustness of the results at all these steps is addressed, to ensure overall robustness and to reduce forecasting *uncertainty*. The methodology allows to build a robust model (stable to changes on the available data), where the error is distributed along the recruitment levels.

A short lived species (anchovy) and medium life-span species (hake, which is suspected to be a predator of anchovy) in the Bay of Biscay are used as application examples. Two interval recruitment discretizations accomplish 70% *accuracy* and *Brier score* of around 0.20 for both anchovy and hake recruitment. In comparison, three intervals recruitment discretizations accomplish 50% *accuracy* and *Brier scores* of around 0.30 for anchovy recruitment, but 0.40 for hake recruitment. These statistics are the result of validating not only the classifier, but also the previous steps, as a whole methodology.

C. Multiple fish species recruitment forecasting by means of multi-dimensional classification methods

A multi-species approach to fisheries management requires a full understanding of the interactions between species in order to improve recruitment forecasting of each of these interrelated species (Hollowed et al., 2000; Edwards et al., 2004). Recent advances in *Bayesian network classifiers* aim the learning of classification models where there are several correlated target variables to be forecasted simultaneously (van der Gaag and de Waal, 2006; de Waal and van der Gaag, 2007). These are known as *multi-dimensional Bayesian network classifiers* (MDBNs).

Pre-processing steps are critical for the posterior learning of the model in domains with sparse and noisy data, such as recruitment forecasting. In the present study, a set of 'state-of-the-art' uni-dimensional pre-processing methods, within the categories of missing data imputation, feature discretization and feature subset selection, are adapted to be used with MDBNs for multi-dimensional domains.

A framework that includes the proposed multi-dimensional supervised preprocessing methods, coupled with a MDBN classifier, is tested on synthetic

datasets. This allows to identify by means of a meta-learning process not only the methods with superior behaviour, but also the circumstances under which a method can have a superior behaviour. Finally, proposals and adaptations of performance measures for the multi-dimensional approach are performed for the measures *accuracy* and *Brier score*.

The conclusions reached are used to apply the approach in recruitment forecasting of several fish species in the ecosystem of the Bay of Biscay for fisheries management. The results show how this approach allows to improve not only the forecasts of each species, but also the forecast of all the species simultaneously.

1.2 Overview of the dissertation

This dissertation is divided into seven chapters, which are organized into three main parts: I) Introduction; II) advanced applications of *supervised classification* in marine science; and, III) conclusions and future work. The first part consists of three chapters. The first chapter is an introduction to the dissertation where the reader can find a synthesis of the contributions and how the dissertation is structured. Chapter 2 introduces biological domains of high interest (mainly for fisheries management) where *supervised classification* can be applied. The main challenges from both points of view, data analysis and biological implications, are described. In addition, biological concepts that are used throughout the dissertation and how the disciplines of marine science and machine learning interact among them are presented. Finally, Chapter 3 is devoted to explaining the classification tasks in machine learning that are later applied in Part II, focussing on *supervised classification*, data pre-processing and model building validation.

Part II is dedicated to the application of *supervised classification* methods to marine science problems related with fisheries management and how these problems have been formulated considering the particular characteristics of each domain:

- Chapter 4 proposes a wrapper method for helping experts in deciding the number of classes or taxa in zooplankton classification.
- Chapter 5 presents a *supervised classification* application to single fish species recruitment forecasting. In this study, a whole robust methodology is proposed, which includes recruitment levels definition, data pre-processing, learning a classifier and honest validation.
- In Chapter 6, the simultaneous multiple fish species recruitment forecasting by means of the multi-dimensional classification approach is presented. In this chapter, a set of 'state-of-the-art' uni-dimensional pre-processing

methods, within the categories of missing data imputation, feature discretization and feature subset selection, are adapted to be used for multi-dimensional classifiers. These proposed methods are tested with synthetic datasets and in a set of real domains of fish recruitment.

Part III concludes the dissertation with Chapter 7. This chapter presents some general conclusions, the list of publications and proposals for future work.

2

Data analysis challenges in marine science

This dissertation lays its foundation in the crossroads between computer science and ecology, particularly in the area of marine ecology and fisheries management. A bridge between both sciences, where they meet together and collaborate (Irigoien, 2006).

In this chapter, a range of problems common in marine science which can be approached by means of supervised classification are presented. This is not an exhaustive list since computer science and, more specifically, *supervised classification* could be applied in almost any domain where data is gathered. The focus of this review is on some problems that are related to fish ecology, as well as fisheries and ecosystems management.

In next section, an introduction to current fisheries management is provided, with examples of the Bay of Biscay. In following sections, a set of common activities that involve data analysis in marine science with fisheries and ecosystems management purposes are presented. These activities have been identified during the author's PhD in the context of the work that is performed in the marine science research facility AZTI-Tecnalia (*http://www.azti.es*). Finally, in the last section a general overview of the general approaches for analysing data in marine science is presented. These approaches can be compared with the *supervised* classification approaches presented in next chapter.

2.1 Fisheries management

The main objective of fisheries management is to accomplish a sustainable exploitation combining long term protection of the resource and economic benefit. Actually, modern fisheries management is often referred to as a governmental system of appropriate management rules based on defined objectives and a mix of management means to implement the rules, which are put in place by a system of monitoring control and surveillance (e.g. 2.1).

As an example, Figure 2.1 shows the flow of advice in TAC (Total Allowable Catch) establishment for some fish stocks in Europe. The stock is esti-

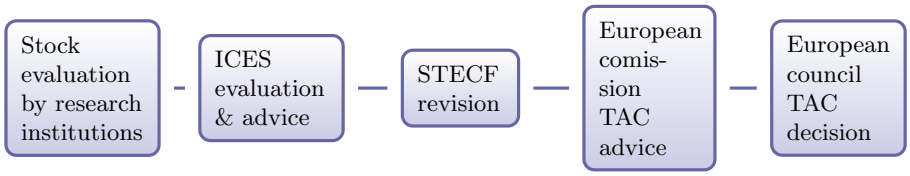


Fig. 2.1. Diagram of current flow of information and advice in European countries for fisheries management decision making.

mated by research institutions, which is evaluated by the International Council for the Exploration of the Sea (ICES), taking into account mainly biological considerations. The ICES working groups provide their recommendations for revision the Scientific, Technical and Economic Committee for Fisheries (STECF) of the European Commission. STECF usually revise the advice considering additional socio-economic considerations providing the TAC advice to European Union Commission. Finally, European fisheries ministers council decides the TAC to be enforced (Villasante et al., 2010).

In practice, the evaluation of stock is an estimate of the actual population. However, fisheries advice has to be given on the projected answer of the population to the fishing pressure. This pressure, together with the actual population, needs to consider new fish entering the fishery (recruitment) and natural mortality (Fig. 2.2) in order to give management advice.

In the actual management system, the 'knowledge base' is oriented towards resource sustainability. Key scientific organisations give formal scientific advice with no economic or social considerations included (Fig. 2.2). However, in its later stages, the decision-making process tends to be highly influenced by economic, social and political considerations (Fig. 2.1). In practice, this has meant that, although long-term sustainability has always been an objective of managers worldwide, a focus on yield optimisation and short-term considerations have determined decision-making outcomes, leading to overexploitation (Worm et al., 2009). In addition, the knowledge base within this system is opaque to the industry and other stakeholders: they do not participate in it and the interpretation of outcomes is divergent among the different participants. All this reduces the reliability of the system, weakens the legitimacy of scientific advice and of the management system as a whole, and increases non-compliance with regulations, which, in turn, has a deteriorating effect on the quality of the information base for the generation of knowledge (Motos and Wilson, 2006).

Nevertheless, research institutions carry out intensive research in order to give better scientific advice. Within this context, it is widely accepted that the exploitation of the sea, together with environmental change, is causing substantial alteration in the marine ecosystem (Anderson et al., 2008). These changes are taking place at a pace that overcomes the ability to manage living resources, as they challenge the capacity of scientists to generate the

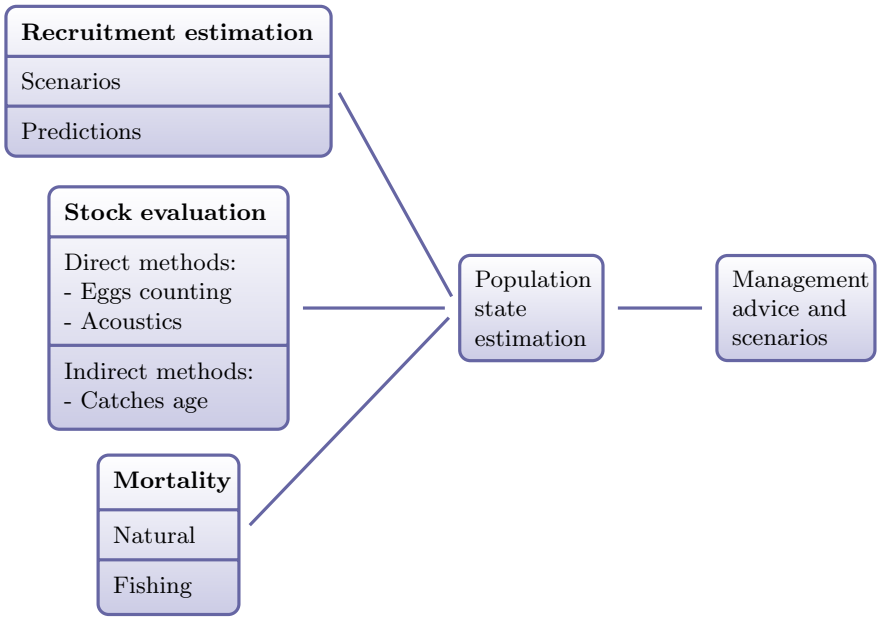


Fig. 2.2. A simplified diagram of knowledge and information acquisition flow for stock estimation with fisheries management advice purposes.

necessary knowledge for effective management. On a global level, the focus on objectives relating to sustainability, long-term yield and maintenance of ecosystem health are widely accepted worldwide.

Indeed, there are many other relationships that have to be understood in order to improve the estimation and advice (Fig. 2.1 base on (Maury, 2010)) for accomplishing the needed multidisciplinary perspective of fisheries management (Motos and Wilson, 2006). The spatial and temporal scales of ocean and fish processes are different and their relationship has to be understood. Nutrients, primary production (plankton), currents and fish are linked in complex ways. Species interactions, especially predation, are also difficult to evaluate. Climate change and its effects on marine systems is an issue of major concern that needs to be addressed.

Therefore, the study of all these factors is the aim of the so-called fisheries oceanography, which aims to explain the super-abundance and collapse of certain fish species (Harrison and Parsons, 2000). Within this context, this dissertation aims to provide methodologies based on the *supervised classification* paradigm, which can help to resolve these needs of knowledge and information with fisheries management purposes.

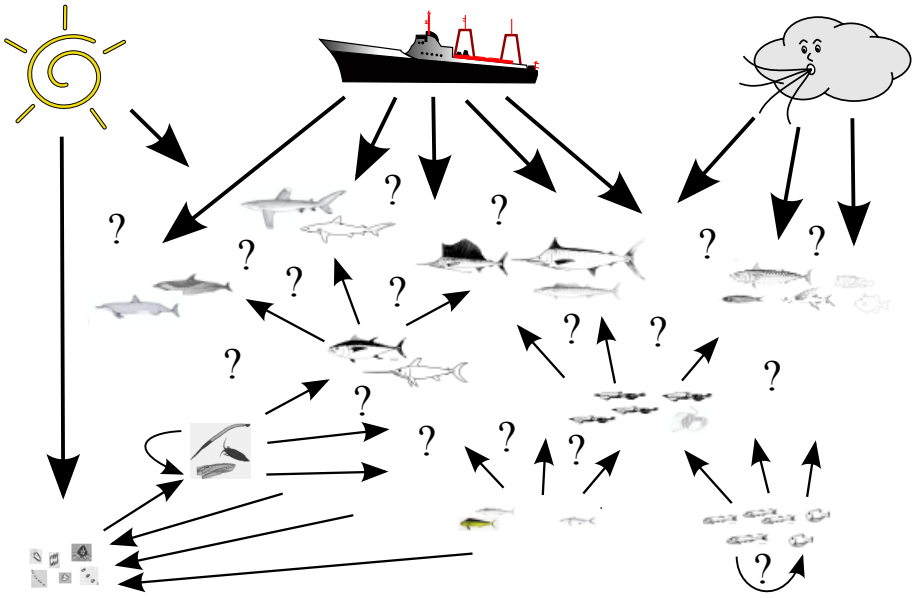


Fig. 2.3. Network of relationships in marine ecosystems. The figure illustrates the effect of atmospheric forcing, both in terms of modifying planktonic production and food fields for fish, as well as the effect of fisheries modifying the trophic links between species.

2.1.1 Ecosystem-based management

Ecosystem-based management is an environmental management approach that recognizes the full array of interactions within an ecosystem, including humans, rather than considering single issues, species, or ecosystem services in isolation (Christensen et al., 1996; McLeod and Leslie, 2009). The traditional approach to fisheries science and management has been to focus on a single species. This can be contrasted with the ecosystem-based approach where other species and the environment are considered.

2.1.2 Water quality directives

On October 23rd 2000, the European Parliament and the Council of the European Union approved the Directive 2000/60/CE, commonly known as the Water Framework Directive (WFD). The WFD constitutes a milestone in the history of environmental policies in Europe, as it changed the way in which the quality of aquatic systems was being monitored and regulated. For the first time, the management measures have a marked marine focus and an integrative point of view. The WFD is important for fisheries management because it goes a step further from ecosystem-based fisheries management.

Ecosystem-based fisheries management focuses on considering the ecosystem of commercial fish species, whereas the WFD has a broader scope of managing the 'health' of full ecosystems.

The implementation of the WFD sets up a challenge in research and environmental policies to all the Member States. The final objective of the WFD is to achieve at least 'good ecological status' by 2015 in all European water bodies. This 'ecological status' has to be assessed based on biological, physico-chemical and hydromorphological elements. Two of the biological elements addressed by the WFD are benthic macroinvertebrates and macroalgae. Each Member State has to select existing tools or develop new ones for the assessment of each of the elements considered by the WFD. Within this context the use of supervised classification methods can be useful to help experts to develop tools that allow them to establish the 'rules' or 'parameters' for managing and assessing the quality of water.

As an example, in 2000, AZTI-Tecnalia (Borja et al., 2000) developed a new tool, based on soft-bottom macrobenthic communities, for the marine environmental quality assessment. This tool, named AMBI (AZTI's Marine Biotic Index), offers a 'pollution or disturbance classification' of a particular site, representing the benthic community 'health' (Grall and Glémarec, 1997). The AMBI is based on previous ecological models, such as those of Glémarec and Hily (1981) and Hily (1984). The theoretical basis is that of the ecological adaptive strategies (MacArthur and Wilson, 1967; Pianka, 1970; Gray et al., 1979) and the ecological succession in stressed environments (Bellan, 1967; Pearson and Rosenberg, 1978). Hence, the species are classified into five ecological groups (EG). The most novel contribution of the AMBI was the formula permitting the derivation of a series of continuous values (Borja et al., 2000).

2.2 Activities and data domains in fisheries research

One of the limitations being faced by the ecosystem approach is the need for additional information. The need for information involves different activities to provide the required data as well as transform it into information and knowledge. These activities include different domains where *supervised classification* methods can be applied. In this chapter, a set of this type of domains is introduced. The domains where the author has accomplished contributions are described in more detail (mainly plankton classification and fish recruitment forecasting); whereas, in the rest of the domains only a short description is provided. The described domains have been grouped in one of four general activities, which correspond with the two subsections of this section. This taxonomy is based on the author's work during the last years of the PhD research in AZTI-Tecnalia:

- Samples classification: Massive amounts of different kinds of samples have to be processed and classified in a short period of time. The most recent approaches combine *image analysis* and *supervised classification* methods.
- Fish recruitment forecasting modelling: to be able to forecast the new fish entering the fishery (recruitment) is crucial to be able to perform resources management.

2.2.1 Samples classification

During oceanographic surveys, most information on physical processes is provided by instruments which allow quick acquisition and storage of data, and subsequently a rapid analysis of the main physical variables. In contrast, chemical and biological studies normally require a collection of water samples, processing them and, in most cases, further analysis in the laboratory. This samples processing is crucial in order to have the necessary information for fisheries management advice (Fig. 2.4).

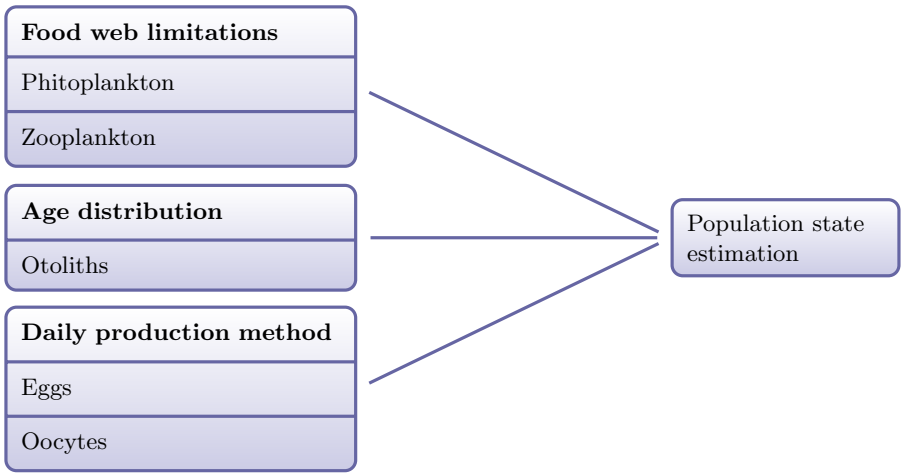


Fig. 2.4. A simplified diagram of samples processing needs for fish stock estimation with fisheries management advice purposes.

This situation often means that there is a long lag between physical and biological oceanographic results and consequently, a delay in obtaining integrated conclusions. In addition, these samples are frequently not analysed completely and many remain unprocessed due to the economic and human cost and the time required to extract useful data (Boyra and Arregi, 2005). Therefore, there is the need to classify massive amounts of biological samples with limited resources. These samples belong to a wide range of types such as *gonad oocyte*, *fish eggs*, *otolites*, *phytoplankton* or *zooplankton* among others.

2.2.1.1 Zooplankton

Zooplankton plays a key role in the transference of primary production (phytoplancton) to fish. Therefore, the study of zooplankton abundance and biomass distribution is important in order to understand marine ecosystems (Irigoien et al., 2002, 2004). Although a routine activity in many laboratories, the amount of samples still presents a practical challenge to marine scientists.

Furthermore, the temporal and spatial sampling scales required to understand the zooplankton distribution (Mackas, 1984; Steele, 1989) are incompatible with the laborious sample analysis using a microscope. To some extent the lack of sample analysis capability has been resolved using simplified measurements such as Chl*a*, total biovolume, biomass or more sophisticated systems providing size and number of particles (e.g. the Optical Plankton Counter). However, all these methods have a common problem: they lack the ability to distinguish between different functional groups of plankton that are known to have a very different role in the ecosystem (e.g. diatoms vs flagellates, marine snow, or copepods vs appendicularia). It is becoming obvious that even proper carbon flux modelisation requires of information on functional groups that is not provided by bulk measurements (Le Quéré et al., 2005).

Therefore, sample analysis is crucial for the understanding of links between fish and ecosystem productivity, in particular plankton samples (Shumway, 1990; Legendre et al., 1991; Longhurst, 1991; Banse, 1995). As an example of one of the difficulties to define the spawning habitat of small pelagics or to understand the variations on biomass (regime shifts) is the lack of appropriate biological information on the prey field for the adult fishes and their offspring. When the physical proxies (temperature and salinity) usually measured during the surveys fail to properly forecast the high production areas, the ability to understand the choice of the spawning habitat is severely limited. Even when the physical proxies are appropriate, they only forecast primary production, which is not necessarily a good factor for the prey field of zooplanktivorous fish. As mentioned earlier, this problem is not restricted to fisheries: knowledge of the factors affecting the distribution of zooplankton is very limited because the difficulties to sample zooplankton with the relevant spatial (mesoscales) and taxonomical resolutions.

Traditionally, plankton have been collected from water samples by filtering or using nets (Wiebe and Benfield, 2003). Zooplankton samples are often digitalized for permanent storage in this format, reducing the storage space needed for conventional plankton samples and preventing the possibility of loss of samples due to the deterioration in the preservative (Alcaraz et al., 2003).

In recent years several *in situ* and laboratory imaging systems have been developed. These systems are capable of obtaining relatively good resolution images at high collection rates that would, in theory, allow to quantify the distribution of taxonomically well resolved groups in the appropriate spatial and temporal scales. These imaging systems (Fig. 2.5) can be classified in

the categories of *scanner based* (Samson et al., 2001; Grosjean et al., 2004; Irigoien et al., 2005), *photographic camera based* (Sieracki et al., 1998; Luo et al., 2005; Bachiller et al., 2010), *video camera based systems* (Davis et al., 1992, 2005; Olson and Sosik, 2007) and *holographic systems* (Loomis et al., 2007; Dominguez-Caballero et al., 2007; Davis, 2008). Exhaustive reviews of such systems can be consulted in Culverhouse et al. (2006); Benfield et al. (2007); Sieracki et al. (2010).

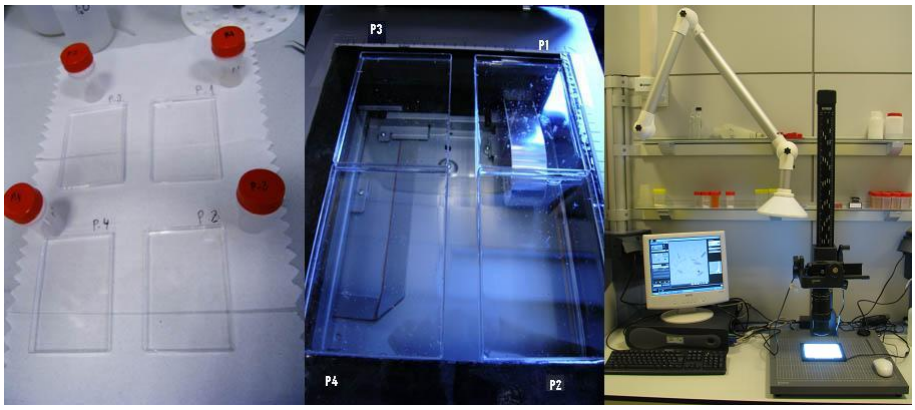


Fig. 2.5. Examples of laboratory imaging systems for zooplankton digitalization. From left to right: samples prepared to be digitalized in plastic cells; simple scanner based system composed of a scanner and a computer; and, photographic camera based system composed of a camera, a robust support with metrics, a led ilumination system and a computer.

These systems have been confronted with a new problem, the huge amount of information (images) produced, which is impossible to analyse manually. *Image analysis* offers an advantage over other methods of counting or sizing: the images can be used for automated species identification using different recognition systems to identify major groups at least (Gaston and O'Neill, 2004; Grosjean et al., 2004). Some of those have been applied to zooplankton with success (Culverhouse et al., 1996, 2003; Grosjean et al., 2004). Furthermore, the monitoring of plankton in this way allows analysis without physical contact, avoiding any likelihood of damage to fragile plankton organisms, such as gelatinous zooplankton (Benfield et al., 2007).

In the case of zooplankton, samples can be stained before scanning for 24 hours with Eosin or Rose Bengal. This process enhances contrast since the former dye stains cytoplasm and muscle protein selectively and the latter has an affinity for lipids (Sheehan and Hrapchak, 1980; Boyra and Arregi, 2005). This staining process can reduce the number of artifact particles between 50% and 75% (unpublished data for Bachiller (2008)), facilitating the task of categorizing the different zooplankton taxa (Fig. 2.6).

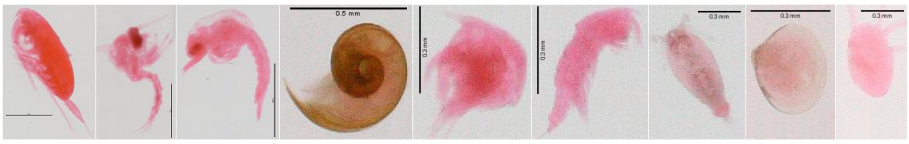


Fig. 2.6. Example of several zooplankton species digitalized by a digital photography camera. From left to right biggest species to smallest: Copepoda Calanus sp., Zoea larva, Furcilia larva, Gastropod, Cirripedia nauplius, Euterpina acutifrons, Bivalve, Oncaea sp., Euphausiid nauplius.

Previous to the counting and classification, the *Image Analysis* process discriminates each object included in the sample from the background by the binarization of the original image (Fig. 2.7). The application of grey levels and colour thresholds determines the minimum levels for a pixel to be considered as a candidate particle for counting. Threshold values must be revised depending on each device and the software used. The minimum and maximum area (expressed as number of pixels) which should be considered to be processed, as well as the *pixels* per *mm* calibration, should be also indicated. Therefore, the establishment of correct values for these parameters is important to obtain satisfactory results. It is also recommended avoiding any pre-processing of images in the digitalization of the samples in order to work with the raw image.

Once the binary images have been obtained, a fixed series of characteristics related to morphological attributes of each particle in the sample is measured, such as the perimeter, area, maximum and minimum diameter and the Equivalent Circular Diameter (ECD). The implementation of different algorithms on these attributes allows other secondary characteristics to be calculated, including the roundness, fractal dimension, elongation and compactness (Fernandes et al., 2009c).

After the different particles have been separated into unique images (vignettes), experts classify a subset of those vignettes of organisms into classes that can be morphotypes and/or taxonomic categories. This set of expert-classified images forms a training set against which classification algorithms can be developed and tested. A full classification framework (e.g. Zooimage: *http://www.sciviews.org/zooimage/*) must include a number of elements: the training set; image analysis methods such as image correction, segmentation and feature extraction (Liu and Motoda, 1998); and a classification algorithm, such as neural network, support vector machine, or decision tree; or an ensemble of algorithms (Sieracki et al., 2010).

The state-of-the art reveals that for automated image classifiers from 10 to 30 classes, the recognition accuracy is over 70% (Blaschko et al., 2005). This is approaching the level of agreement among human experts (Culverhouse et al., 2003; Bell and Hopcroft, 2008; Gislason and Silva, 2009). Bias due to errors in classification can be statistically corrected if the prior probabilities of the occurrences of the types are known (Solow et al., 2001). A carefully collected

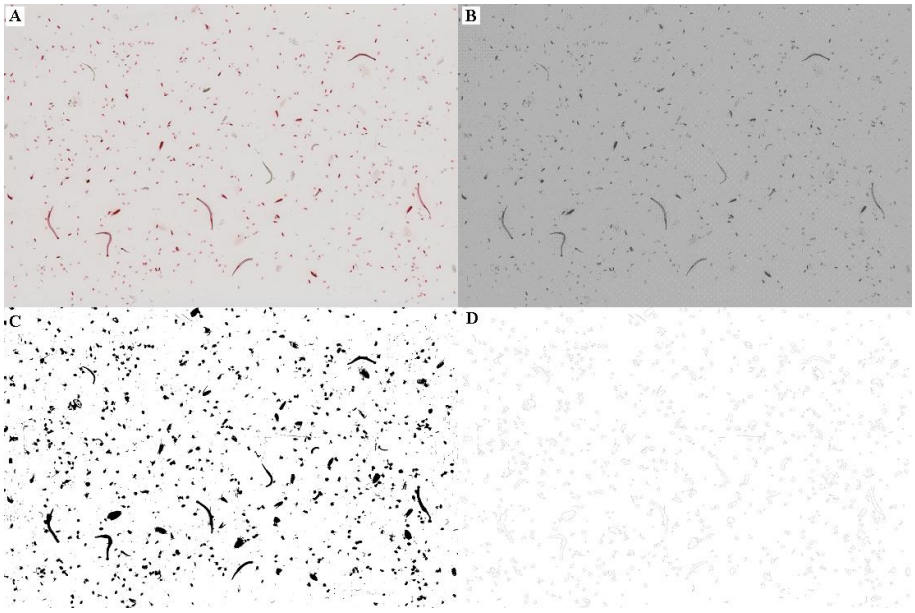


Fig. 2.7. Image analysis process in zooplankton samples: A) original color image; B) image transformed to gray scale; C) image binarized for silhouette identification; and, D) silhouette of images to produce measurements of features

expert-derived training set can provide these prior probabilities. Misclassification may also be reduced by considering results from multiple classifier approaches (Hu and Davis, 2006), or optimizing class selection (Fernandes et al., 2009c). However, more work on handling the errors in classification, and on tools and protocols for creating appropriate and unbiased training sets is needed.

An important issue is the availability of public datasets. Indeed, a classifier cannot be better than the dataset used to train it (Irigoien et al., 2005). These datasets must be provided by the taxonomists, but statistical help is required to establish procedures that both quantify and minimize inevitable errors.

Continued work to identify features and create improved classification algorithms is needed in this field. It has been suggested that a community effort of open source software development is the best way to make progress in this area (RAPID: Research of Automated Plankton Identification (Benfield et al., 2007)). Examples of such software development are the Plankton Analysis System (PAS) and the Plankton Interactive Classification Tool (PICT) being developed at the University of Massachusetts Amherst (Mattar et al., 2009). PAS is a web-application that provides the functionality for experts to upload their images and algorithms, process images, hand-label exemplars, train classifiers and use those classifiers to automatically label new images. Zoo/PhytoImage software has been successfully employed in a number of stud-

ies (Zarauz et al., 2008; Bell and Hopcroft, 2008; Irigoien et al., 2009) as tools for automatic identification of scanned meso and macrozooplankton images.

Finally, another problem is that usually the datasets are imbalanced (Japkowicz and Stephen, 2002). The class imbalance problem corresponds to classification domains for which one class is represented by a notably larger number of instances than other classes. A common practice for dealing with imbalanced data sets is to rebalance them artificially (Provost, 2000). This has been called 'upsampling' (replicating cases from the minority) (Zarauz et al., 2008) and 'downsampling' (ignoring cases from the majority) (Grosjean et al., 2004). However, in some domains it has been demonstrated that upsampling or downsampling does not solve the problem (Provost, 2000; Drummond and Holte, 2000). The reason is that these techniques can show an improvement of the estimated performance that might not be real. This is difficult to verify since the real distribution at sea is unknown. One approach to validate estimations from automatic classification is comparing them with manual classification and counting (Bell and Hopcroft, 2008), which also has a degree of error (Culverhouse et al., 2003), or with some type of control artificially introduced in the samples such as control beads (Bachiller et al., 2010).

In conclusion, there has been a lot of effort in improving the devices to acquire more and better quality data. However there is the need to improve the posterior steps of data analysing in order to get the most from it and to ensure that the conclusions extracted are valid.

2.2.1.2 Phytoplankton

Traditional phytoplankton surveys have been limited by the laborious and time-consuming nature of sampling and analysis similarly to the zooplankton problematic. Such a problem of sampling planktonic organisms at the relevant scales has hampered the understanding of control mechanisms in marine systems (Duarte, 2007).

The field of phytoplankton research has benefited from the presence of photosynthetic pigments, which has permitted the use of colour and fluorescence to study the distribution with high spatial resolution (by means of satellite imaging methods, in situ fluorescence measurements, etc), and with some taxonomic differentiation methods (high performance liquid chromatography). However, to fully understand the ecosystem, information is needed on the distribution of prey and predators: relative abundances, size distribution, overlapping, etc.

Until recently, obtaining high-resolution distribution of heterotrophs regularly at sea was not possible. However, within the last few decades, new image analysis systems have been developed for rapid and high-resolution data acquisition. These systems are designed for studying a broad range of sizes (Ashjian et al., 2001; Davis et al., 2004; See et al., 2005) and constitute a powerful alternative to the traditional manual treatment of plankton samples.

At present, automatic sampling methods still lack taxonomic detail (Hu and Davis, 2006). Nevertheless, recent studies have demonstrated the power of machine learning and data mining technique used to classify field-collected organisms of different taxa (Culverhouse et al., 2003; Blaschko et al., 2005) in achieving good accuracy levels in terms of abundance and biomass of major taxonomic groups (Culverhouse et al., 2003, 2006). These systems do not have the resolving power to identify plankton to the level of species and life stage, but can provide important information on coarse taxonomic composition (Davis et al., 2005). Recent studies start to accomplish taxonomic composition by means of *image analysis* and *supervised classification* (Zarauz et al., 2009).

2.2.2 Fish recruitment forecasting

Early on in fisheries research, recruitment was identified as a key element in management, i.e. the amount of fish that enters the population each year. As a result, recruitment and the factors determining it have been the subject of intense research (Ricker, 1954; Cushing, 1982; Myers et al., 1995). A problem is that data about some of the factors that can be controlling recruitment directly (e.g. food availability, larval growth) may be more laborious to obtain than the recruitment estimate itself. In addition, the interactions between population dynamics and different environmental factors are complex and often non-linear, making it difficult to produce robust predictions. Within this context, recruitment forecast is a problem of high *uncertainty* (Mantyniemi et al., 2009).

Based on a simplified approach, fisheries management has been moving towards the use of environmental relationships using oceanographic data, since environmental factors are collected routinely (Bartolino et al., 2008; Borja et al., 2008). Therefore, such research has evolved from considering only the biomass of spawners, to also including environmental factors that can modulate recruitment (Schirripa and Colbert, 2006; Planque and Buffaz, 2008). Therefore, it is well accepted that environmental conditions and climate play an important role in the recruitment of fish (Cushing, 1982; Baumgartner et al., 1992; Bakun, 1996; Alheit and Hagen, 1997; Brunel and Boucher, 2007; Borja et al., 2008). In this field, a large number of studies have been undertaken, using different techniques, to consider such environmental information to forecast recruitment (Chen and Ware, 1999; Bailey et al., 2005; Dreyfus-León and Chen, 2007; Dreyfus-León and Schweigert, 2008; MacKenzie et al., 2008; Ruiz et al., 2009).

Nevertheless, the problem remains difficult because the mechanisms behind such relationships are often poorly understood; this, in turn, makes it difficult to determine the forecast estimation robustness, leading to the failure of some proposed relationships, methods and performance estimations, when new data become available (Myers et al., 1995). Such failures may be related to new controls, which were not considered previously (Myers et al., 1995;

Planque and Buffaz, 2008), or to limitations in the available data (Schirripa and Colbert, 2006). The main limitation to achieve good forecasts, from a data analysis perspective, is the sparse and 'noisy' nature of the available data (Francis, 2006) and the need to estimate the risk or *uncertainty* of that prediction. Probabilistic classification models offer the possibility to establish the *uncertainty* associated with a prediction (Friedman et al., 1997), in addition to the model performance estimation.

Many studies have addressed recruitment forecasting by selecting the features and learning a model for each species in isolation. Most mono-species approaches have used regression methods (Ricker, 1954; Beverton and Holt, 1957; Schirripa and Colbert, 2006; Planque and Buffaz, 2008) and Bayesian statistics (McAllister and Ianelli, 1997; Meyer and Millar, 1999; Newman et al., 2006; Ibaibarriaga et al., 2008). The importance of modelling entire ecosystems, rather than single species, is leading to multi-species management approaches that take into account interactions between species (Edwards et al., 2004; Fernandes et al., 2010b). Some of these interactions are due to competition for food and space, as well as predation between species (Ragozin and Brown Jr, 1985; Leggett and Deblois, 1994; Fortier and Villeneuve, 1996). There is a broad set of models of different categories that are based on this multi-species management approach as summarized in (Hollowed et al., 2000). However, these models require an expert to specify all the characteristics and the relationships that are relevant, which are often unknown. This leads to complex models, which are often impossible to parameterise with sparse and noisy data (Botsford et al., 1997; Fulton et al., 2003; Essington, 2001).

Previous *supervised classification* studies for recruitment forecasting have accomplished good forecast rates for Pacific herring recruitment using genetic algorithms to produce prediction rules (Dreyfus-León and Chen, 2007; Dreyfus-León and Schweigert, 2008). However, probabilistic models are not used in these studies; and therefore, predictions are provided without estimations of their likelihood. In addition, these studies assume that experts already know the best set of factors for recruitment prediction. This assumption is not necessarily true for many species; rather, it often happens that predictive environmental variables are proxies for unknown processes, or that new data provide better explanatory variables. In this dissertation, the most appropriate and smallest set of factors is considered to remain unknown, although there is a large group of good 'candidates' proposed by experts (Jin, 2009). Therefore, the 'best' factor set identification is part of the proposed methodology (Saeys et al., 2007).

2.3 Data analysis in marine science

Marine science is a science trying to answer difficult questions with limited resources. In practice, one has to apply a data exploration, check assumptions, validate the models, perhaps apply a series of methods, and most importantly,

interpret the results in terms of the underlying biological questions being investigated and the limitations of the collected data (Zuur et al., 2007).

The biological question aids in deciding on the type of analysis. In addition, the quality of the data (number of variables, how many observations, search for linear or non-linear relationships) decides the specific method, which can only be addressed by a detailed data exploration. Therefore, data exploration is the first step in a data analysis procedure in marine science (Fig. 2.8).

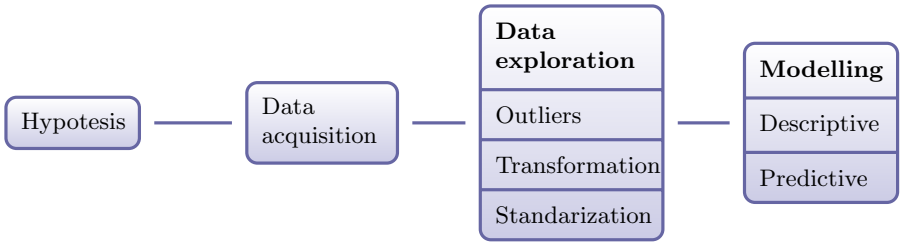


Fig. 2.8. Example of a common pipeline of data analysis in many marine science research works.

The data exploration step allows to decide the appropriate modelling method and whether other pre-processing methods have to be applied. In this sense there is a wide range of data exploration techniques (Montgomery and Peck, 1992; Crawley, 2002; Fox, 2002; Quinn and Keough, 2002), where some of the most useful are (Tague, 2005; Zuur et al., 2007):

- *Boxplots* and variants (Chambers, 1983) allow to find relationships between variables.
- *Dotplots* (Cleveland, 1984) are useful to identify outliers and homogeneity.
- *Histograms* show the centre and distribution of the data and give an indication of normality.
- The *Quantile-Quantile plot* (Wilk and Gnanadesikan, 1968) is a graphical tool used to determine whether the data follows a particular distribution.
- The *Scatterplot* or *scattergraph* allows to detect relationships between pairs of variables.
- The *pairplot*, or *scatterplot matrix* shows multiple pair-wise scatterplots in one graph and can be used to detect relationships between variables and to detect collinearity.
- The *coplot* is a conditional scatterplot showing the relationship between two variables for different values of a third variable or even a fourth variable.
- *Lattice graphs* show relationships between two variables which are conditioned on nominal variables, with the advantage over *coplots* of being able to work with larger numbers of panels. However, they have the disadvantage that the conditional factor must be nominal.

- *Design and interaction plots* visualise differences in mean values of the target variable for different levels of the nominal variables and interactions between attribute variables.

Another important step is *outlier identification*. An outlier is a data point that, because of its extreme value compared to the rest of the dataset, might incorrectly influence an analysis. For example, Principal Component Analysis (PCA) depends on linear relationships between variables, and outliers may cause non-significant regression parameters and mislead the analysis. The outlier identification prompts to check the original data to find errors in data entry or whether there was some data recording failure. However, often there is no data error, just the result of variability (Irigoien, 2006). In these cases, outliers can be often dealt with by applying a *data transformation* such as *squared root transformation*.

Most transformations of the variables do not influence the results when some techniques, such as *classification trees* or *probabilistic graphical models*, are applied. Whereas, for others, such as regression based techniques, applying the proper transformation can be crucial. The transformation to be applied is often selected based on the experience of data analysts or using the Mosteller and Tukey's bulging rule (Mosteller and Tukey, 1968; Fox, 2002) or by automatic transformation selection (Montgomery and Peck, 1992; Fox, 2002). Similarly, if the variables being compared are from widely different scales, such as comparing the growth rates of small fish species against large fish species, then standardisation (converting all variables to the same scale) might be necessary depending on the modelling approach selected (Zuur et al., 2007).

Once the data is prepared, often regression is applied, particularly, if there is sparse data or as an approach to describe the data with continuous or discrete target variables. However, sometimes there is no target variable and the goal is to identify relationships in data. If there is enough data and the relationships are really linear it is also possible to make predictions using basic regression models. However, there are often complex non-linear relationships between physical and biological processes. In the following paragraphs a general overview of different modelling approaches used in marine science and environmental modelling are summarized (Guisan and Zimmermann, 2000; O'Brien et al., 2004; O'Brien, 2004; Guisan and Thuiller, 2005; **?**).

New statistical methods have been developed to model spatial and temporal dependence based on regression such as *generalized linear models (GLMs)* or *generalized additive models (GAMs)*. GLMs and GAMs were first developed in the 60s and have since been used extensively in marine science research (Guisan et al., 2002; Guisan and Thuiller, 2005). GLMs are extensions of linear models, allowing for non-linearity and non-constant variance structures in data. GAMs (Hastie and Tibshirani, 1990) are a further extension of GLMs, where the only underlying assumption is that the functions are additive and the components are smooth (Guisan et al., 2002; Zuur et al., 2007).

Through the use of spline functions, GAMs allow the data to accommodate the shape of the response curves to almost any functional form. These models are particularly useful in marine modelling because underlying data is usually highly nonlinear and may take on many different distribution forms. Moreover, GAMs are able to handle multi-collinearity between variables (Yee and Mitchell, 1991), and to minimize the effects of extreme observations (Wood, 2008). Because of all this, GAMs have demonstrated a great ability to model complex non-linear relationships between variables, and so have been applied in many fields of marine research, both for terrestrial (Guisan and Zimmermann, 2000) and marine systems (Augustin et al., 1998; Beare et al., 2000; Stratoudakis et al., 2003; Planque et al., 2007).

These methods are well suited to marine presence/absence (or true/false) data, in particular where little is known about the relationship between factors and species presence. However, in essence, GLM and GAM are extensions of regression and therefore face many of the same issues regarding their applicability, mainly in domains where it is necessary to manage *uncertainty*.

Logistic regression (Hosmer and Lemeshow, 1989) is another popular technique in habitat modelling and is capable of producing good results, providing a number of assumptions are met. Logistic regression is useful for predicting a binary response from either continuous or categorical factors. Although logistic regression is a robust method, it needs relatively large datasets.

Artificial neural networks (ANNs) (McCulloch and Pitts, 1943b) are an artificial intelligence technique based on a representation of the neural interactions in the human brain. Information is passed through a number of nodes, resulting in values or classifications. In domains with sparseness of data and the requirement of being able to incorporate expert knowledge, ANN is a less promising method. In addition, in most domains of marine science the model must display low structural complexity and must be easy to communicate and to implement spatially (O'Brien et al., 2004), which often does not suit the black box approach of ANNs.

In *classification and regression decision trees (CART)* (Breiman et al., 1984) a decision is taken at each node in the tree depending on the observation value, and with leaves of the tree representing resulting classifications. The amount of data required to specify robust trees is a drawback in the case of many marine science applications. However, the fact that expert opinion can be incorporated relatively easily is beneficial. In addition, trees can be easily interpreted for biological meaning. It is not clear, however, how to deal effectively with *uncertainty* in decision trees. CART could be a useful tool for organising data and incorporating expert knowledge.

Fuzzy logic (Zadeh, 1965) seeks to relax the crisp and deterministic classifications imposed by Boolean logic. Fuzzy membership generalises Boolean logic by assigning the value 1 to the state 'true', 0 to the state 'false' and allowing values between these two numbers. The many uncertainties, both in geographical and attribute space, could be addressed using fuzzy classifi-

cation. This allows for classifications of 'marginally suitable', in addition to classifications of 'not suitable and 'suitable'.

Often in marine science both, *Bayesian statistics* (Laplace, 1912) and *Bayesian networks* (Pearl, 1985), are called *Bayesian methods*. Both paradigms are based on the *Bayes rule*. However, *Bayesian networks* do not necessarily imply a commitment to *Bayesian statistics*. Indeed, it is common to use *frequentists methods* to estimate their parameters (although *Bayesian statistics* can be also used). Bayesian methods provide a 'formalism for reasoning under conditions of *uncertainty*, with degrees of belief coded as numerical parameters, which are then combined according to rules of probability theory' (Pearl, 1988). A simple Bayesian model defines prior and conditional probability distributions for each node and then uses combination rules to propagate conditional probability distributions through the network. The probability distributions may be derived from data, set by experts or defined from a combination of data and expert opinion. This process of combining probabilities produces conditional probabilities for each possible outcome.

Most statistical techniques are unworkable in this situation, with too many assumptions needing to be made in order to perform any analysis. Bayesian modelling techniques provide a simple yet robust way of managing *uncertainty* explicitly in the form of probabilities. The advantages of *Bayesian Networks* over GLMs or GAMs applied to marine modelling can be observed in O'Brien et al. (2004).

As an example of the applications of those techniques the book of Zuur et al. (2007) can be consulted. The works Guisan and Zimmermann (2000) or Guisan et al. (2002) are also a good general introduction in the application of those techniques to different biological domains and they provide comparison with other modelling approaches. Some other recent applications to the field of plankton are Zarauz et al. (2007) and Zarauz et al. (2008). In the field of fish species distribution Planque et al. (2007) can be consulted.

In some scientific fields, it is the general belief that it is needed to formulate a hypothesis in advance and specify every step of the statistical analysis before starting. Although it is agreed that there must be always at least some intuition behind an analysis, deciding on the statistical methods before seeing the data is unrealistic in most environmental studies (Graham et al., 2004; Zuur et al., 2007). Even in the early stages of a marine experiment, survey or monitoring programme, it is highly likely that the generated data are so noisy that the pre-specified method ends up unsuitable, forcing the exploration of other alternatives. Another vision has no initial hypothesis formulated and the space of possible hypothesis is explored, as is performed in data mining approaches. However, there is often sparse data available which needs the application of robust machine learning techniques in order to extract valid conclusions.

3

Supervised classification

In the previous section several marine science domains where machine learning can be applied have been presented. Within the context of the presented domains, in this section, several concepts belonging or related with *supervised classification* are presented. However, these introduction of these concepts is limited to what is necessary in order to understand the contributions of the author to these domains in following chapters (see Part II).

3.1 Introduction

Artificial intelligence, machine learning and data mining are branches of computer science closely related between them and with statistics and mathematics. Their aim is to allow computers to perform complex tasks that involve learning or reacting to data in an 'intelligent' way. Three main types of task could be differentiated in machine learning:

1. *Supervised classification*: An expert labels a set of data (training-set) in a limited number of groups (classes or labels). Labelled data (e.g. Table 3.1) is used to learn a model (classifier) in order to classify new unseen data in the defined groups (Duda et al., 2001; Alpaydin, 2004; Bishop, 2006).
2. *Unsupervised classification*: There is no expert labelling provided and groups or labels are created searching for similarities in data by means of automatic methods (Forgy, 1965; Jardine and Sibson, 1971; Dempster et al., 1977; Bezdek, 1981). The hope is to discover unknown, but useful, patterns in data (Jain et al., 1999).
3. *Semi-supervised classification*: In many domains, there is a small amount of labelled data and a large amount of unlabelled data (Blum and Mitchell, 1998; Zhu, 2006; Chapelle et al., 2006; Calvo et al., 2007). This approach aims to take advantage of both, labelled and unlabelled data. The aim is to learn a classification model in domains where it is hard to get labelled data or it is too expensive.

Table 3.1. Example of zooplankton images dataset for *supervised classification*. The first four features are morphological, extracted by means of image analysis: minor, the smallest axis of the ellipsis containing the individual; area, the surface area occupied by the individual; ECD, Equivalent circular diameter; mean of the grey scale of the pixels of the individual image. The next two features are environmental measurements at the collection time of the samples: salinity and temperature. The final column shows the zooplankton taxa of that individual labelled by an expert (class variable).

Minor	Area	ECD	Mean	Salinity	Temperature	Class
0.6282	156.5	0.79	15.4	34.6	16.6	Copepod
0.4797	106.3	0.82	41.6	35.8	17	Copepod
0.4629	101.5	0.7	40.3	35.1	16.5	Euphausid
0.164	171.2	1.51	2.7	35.9	16.2	Artifact
1.3975	96.9	1.59	42.1	34.6	16.6	Decapoda
0.3765	105.5	0.63	38.4	35.3	15.9	Copepoda
0.5354	106.1	0.89	41.4	35.4	15	Copepoda
...	...	...	...	...	...	...

Informally, *supervised classification* can be understood as learning to distinguish concepts from experience (Pérez, 2010), e.g. learning to distinguish different species of plankton from the characteristics extracted from their image (Fernandes et al., 2009c). Usually, the experience is represented by a set of examples (instances, cases, individuals or samples) of the given concepts, e.g. the available collection of characteristics extracted from images of different types of plankton. In this example, a case or instance contains the features, variables or factors extracted from an individual specimen image as well as a special variable that contains the label or assigned classification to that specimen. This target variable is usually called class variable or class. A set of instances (or dataset) permits to learn a classifier, that is a function which assigns a class label to an unlabelled case described in terms of its features.

The objective of *supervised classification* consists of building a classifier from training data S, with N cases (Table 3.2), in order to classify the value of a single target variable or class variable C, given the set of factor or feature variables $\mathbf{X} = (X_1, \cdots, X_n)$ of an unseen unlabelled case $\boldsymbol{x} = (x_1, \cdots, x_n)$. As an example, in fish recruitment, (C) represents recruitment of a fish species to be forecasted and $(X_1, \cdots, X_n)$ represents the set of factors (climatic, biological and others) or features.

In this work, the term 'forecasting' is used instead of 'predicting', since the verb 'to predict' has the connotation of 'guessing', although the terms 'forecasting' or 'diagnosing' would be more appropriate becouse they that have the meaning of 'determining from evidence'. Similarly, the terms 'features' or 'factors' are used instead of terms suchs as 'predictors'.

In order to learn a classification model, it is often necessary to perform a set of previous pre-processing tasks to prepare the data for model learning

Table 3.2. Data matrix of a *supervised classification* problem.

	X_1 ... X_n	C
Instance 1	x_1^1 ... x_n^1	c^1
Instance 2	x_1^2 ... x_n^2	c^2
Instance 3	x_1^3 ... x_n^3	c^3
...	...	
Instance N	x_1^N ... x_n^N	c^N

(pre-processing). These tasks can be crucial in certain domains such as those with sparse data. Some examples are data cleaning, missing data imputation, discretization or feature subset selection. Some machine learning concepts are defined in the following sections in order to understand the contributions of the author in Part II:

- Firstly, the process of data analysis common in *supervised classification* is introduced, which can be compared with the process common in marine science data analysis previously presented (Chapter 2).
- Secondly, the concept of *supervised classification* is introduced.
- Thirdly, filter pre-processing methods are presented. Most of them are based on *uncertainty reduction*, therefore, an introduction to some concepts of information theory is needed. The methods are in the areas of missing data imputation, discretization and feature subset selection.
- Fourthly, *Bayesian network classifiers* are presented as a modelling approach to deal with needs identified in the domains of the previous chapter (high *uncertainty*). The *Markov blanket* concept of a *Bayesian network* is also presented.
- Fifthly, the need for assessing the performance of applying a method, or group of methods (pipeline), and techniques to accomplish it are introduced. In addition, the need to evaluate not only methods alone, but full pipelines of the methods is addressed as well as the issue of the statistical comparison of methods.
- Finally, the previous concepts are presented for the multiple class variables approach (multi-dimensional).

3.2 The process of data mining

The process of data mining consists of several steps, from data collection to learning a final model and the interpretation of results. The first step, the data collection, often implies other areas of expertise. The whole process does not just consists of learning a model from the collected data (Fig. 3.1). In most cases, real-domain collected datasets are often not directly suitable for model

induction, they contain noise and missing feature values, and therefore a significant pre-processing effort is required (Zhang et al., 2003). In particular, in certain domains such as those that manage annual climatic and environmental variables, data are sparse and difficult to manage. In these domains, appropriate data pre-processing can dramatically condition the final model and its performance (Fernandes et al., 2010c; Uusitalo, 2007). In particular, when the pre-processing method uses the value of the *class variable*. They are known as *supervised pre-processing methods* (Dougherty et al., 1995; Kotsiantis et al., 2006). Finally, it is also needed to estimate the goodness of the model building process (pre-processing and classification model) in order to assess its usefulness, power and reliability.

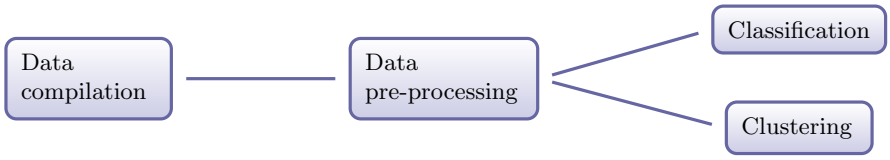


Fig. 3.1. Example of a common data analysis flow.

Within this context, a pipeline that includes an ordered set of pre-processing supervised methods before learning the final classifier is common (Fig. 3.2) in *supervised classification* tasks. A pipeline common in the literature (Dougherty et al., 1995; Kotsiantis et al., 2006; Fernandes et al., 2010c), which is going to be recurrent in this manuscript, is formed by the following steps: 1) missing data imputation; 2) feature discretization; 3) feature subset selection; and 4) classifier learning. However, these and other additional steps can be present, depending on specific necessities and characteristics of each domain.

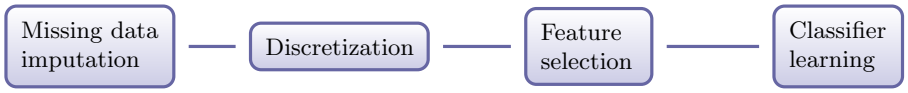


Fig. 3.2. Model building pipeline.

In the next section a formal definition of *supervised classification* is provided. Then, information theory concepts, that are needed to understand later explained pre-processing methods, are described.

3.3 Supervised pre-processing methods

Many *supervised* machine learning models require a previous data pre-processing step. There are algorithms that can not be applied in the presence of missing

data. Other algorithms require discrete data. There is a broad set of pre-processing methods. In this dissertation, the focus is on the missing data imputation of continuous variables, discretization and the feature selection. Pre-processing methods can be classified by taking into account whether or not they use the posterior classification learning algorithm. While wrapper methods make use of the classification algorithm (Kohavi and John, 1997), filter methods (Saeys et al., 2007) use metrics based on intrinsic data characteristics. In addition, filter methods are usually faster than wrapper methods (Amaldi and Kann, 1998; Inza et al., 2000) using metrics such as *uncertainty* reduction (Ben-Bassat, 1982; Fayyad and Irani, 1993; Cover and Thomas, 2006) or correlation measures (Hall, 2000).

Therefore, the use of filter methods for pre-processing tasks is preferred in this dissertation, except in tasks involving the definiton of the class variable where wrapper approaches have been used. In addition, methods based on entropy (*uncertainty*) reduction from information theory are selected due to the high *uncertainty* nature of the domains dealt with in this dissertation. Consequently, in the following sections an introduction to some concepts of information theory is needed, prior to introducing methods in the areas of supervised missing data imputation, discretization and feature subset selection.

3.3.1 Information theory in supervised pre-processing

Information theory (IT) (Cover and Thomas, 2006) is a branch of applied mathematics and electrical engineering involving the quantification of information. IT permits to measure the *uncertainty* and dependence between variables. Those measures allow an appropriate framework for the development of machine learning methods, based on dealing with *uncertainty* (Jakulin, 2005). Therefore, the formulation of several measures from information theory is provided. The formulations are limited to discrete or case of categorical variables, since this is the common use across this dissertation. However, further information for those measures calculation for both type of variables, discrete and continuous, can be consulted in Pérez (2010).

Entropy:

The entropy of a random variable C quantifies its *uncertainty*. It measures the number of bits needed to code the variable. The lower the *uncertainty*, the lower the number of bits needed to code it.

$$H(C) = -\sum_{c} p(c) log_2 p(c)$$

Conditional entropy:

Conditional entropy quantifies the remaining *uncertainty* of a variable C given that another variable X is known:

$$H(C|X) = H(C,X) - H(C)$$
$$H(C,X) = -\sum_{c,x} p(c,x)\log p(c,x)$$

As an example, the difference between the entropy of a variable C and the entropy after an additional variable is provided X (conditional entropy) is behind one popular discretization method (Fayyad and Irani, 1993). This method aims to find cut-off points of X that minimize the *uncertainty* of C (or maximize the dependence between C and X), which is quite intuitive and useful for later expert interpretation. Note: If $H(C|X)$ is 0, it means that there is complete dependence between C and X; and therefore, C can always be forecasted if X is known.

Mutual information (MI):

Mutual information quantifies the mutual dependence of two variables (Shannon and Weaver, 1963) based on the previously explained concepts of *entropy* and *conditional entropy*.

$$MI(C,X) = H(C) - H(C|X)$$

Several variations on mutual information have been proposed to suit various needs such as normalized variants.

Symmetrical uncertainty score (SUS):

SUS is a normalized variant of mutual information that has been used for ranking pairs of variables (Hall, 1999), providing measures of MI between 0 and 1 that are easier to be interpreted by experts than the non-normalized variants.

$$SUS = 2\frac{MI(C,X)}{H(C)+H(Y)}$$

As an example, SUS is behind one popular feature selection method, correlation based feature subset selection (CFS), that has an interesting formulation (Hall and Smith, 1997; Hall, 2000). In the formulation of this method the space of possible feature subsets is explored. Subsets are given a merit where dependences between features and the class are prized, whereas dependences between selected features are penalized, being those dependences measured by means of SUS between pairs of variables. This formulation also provides a high degree of interpretability for experts.

K-way interaction:

Previous concepts measure the *uncertainty* in a variable or pair of variables. However, it is possible to measure a higher order interaction between

more than two variables. This can be done by means of the k-way interaction measure (Jakulin, 2005). Mutual information is a special case when there are only two variables.

$$I(X_1;...;X_k) = -\sum_{\boldsymbol{Y}\subseteq\boldsymbol{X}} (-1)^{k-|\boldsymbol{Y}|} H(\boldsymbol{Y})$$

Figure 3.3, adapted from Pérez et al. (2006), summarizes the relationship of the measures that have been introduced.

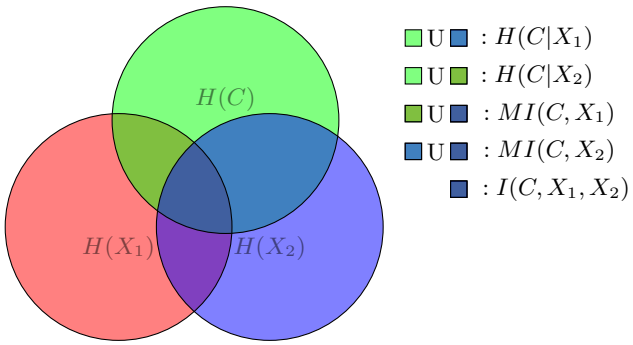


Fig. 3.3. Graphical representation of several information theory *uncertainty* measures

3.3.2 Supervised missing data imputation

In many domains, not all the feature values are known for all the cases (Batista and Monard, 2003). This raises problems in order to apply many methods that can not be applied in the presence of these missing values. There is a broad range of methods that can be applied to impute missing values (Kononenko et al., 1984; Smith et al., 1996; Allison, 2001; Delavallade and Dang, 2007). However, in this dissertation the *CMean* method (CM) has been selected given its simplicity and effectiveness in a broad set of domains (Kononenko et al., 1984; Little and Rubin, 2002; Delavallade and Dang, 2007). In this method, given a missing value in an instance of a feature, it is filled with its mean (continuous variables) or its mode (discrete variables), considering only the instances that present the same class variable label as the instance with the missing value. In this dissertation the missing imputation is performed over continuous variables:

$$x_i^k = \frac{1}{N_c} \sum_{N_c} x_i$$

where i is the feature index and k the instance number of the missing value. The subset of instances where the class label is the same as in the missing instance is denoted by N_c.

3.3.3 Supervised discretization

Supervised discretization of features implies the transformation of continuous variables into categorical variables, taking into account the class values. In some context, it presents several advantages over the use of the original continuous values (Fernandes et al., 2010c): a reduction in time to induce a classifier (Fayyad and Irani, 1993; Dougherty et al., 1995); an enhanced capability to interpret the model outputs and structure (Geurts and Wehenkel, 2000); and an improvement in classification performance (Dougherty et al., 1995; John and Langley, 1995; Blanco et al., 2003).

As a supervised discretization technique, the state-of-the-art Fayyad and Irani's Multi-Interval Discretization (MID) method (Fayyad and Irani, 1993) has been selected. It searches recursively, in each feature, for a set of cut-off points that reduces the class entropy. This method firstly searches for the cut-off point of the given feature X_i that minimizes the conditional entropy $H(C|X_i)$ of the class variable C. In following recursive searches, the method repeats the process on both sides of the previous selected cut-off point. The process is stopped if the *gain* in entropy reduction $H(C) - H(C|X_i)$ is below a Minimum Description Length (MDL) criterion (Rissanen, 1978):

$$gain > \frac{1}{N}\left(\log_2(N-1) + \Delta\right)$$

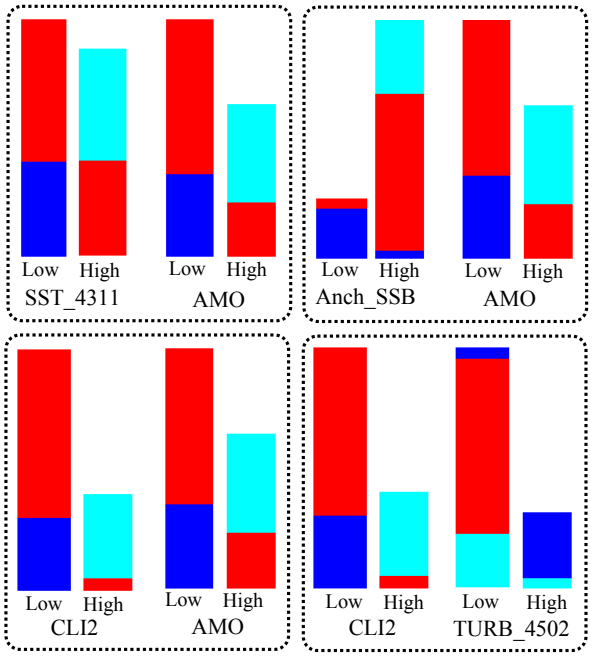
$$\Delta = \log_2(3^r - 2) - [rH(S) - r^1 H(S_{left}) - r^2 H(S_{right})]$$

where r is the number of class values present in the full training data S and r_1, r_2 are the number of class values in each resultant data subset after applying a cut-off point (S_{left}, S_{right}).

3.3.4 Supervised feature subset selection

Feature subset selection (FSS) (Yu and Liu, 2004; Saeys et al., 2007; Guyon et al., 2007) is the process of reducing the number of features before learning a classifier. It has three main advantages; 1) improvement of classifier performance; 2) provision of more cost-effective features; 3) a better understanding of the underlying processes that generated the data.

The popular multivariate Correlation-based Feature subset Selection (CFS) method (Hall and Smith, 1997; Hall, 2000) has been selected in this work as a prior step to classifier learning. CFS is based upon an intuitive formulation, the assumption that a good subset of factors is one that is highly correlated with the class and, at the same time, the features have low correlation between them (Fig. 3.4).



Factor	SUS
CLI2	0.58
Turb_4502	0.47
AMO	0.42
Anch_SSB	0.38
SST_4311	0.37

Fig. 3.4. CFS formulation searches for subsets of variables that complement each other given a target variable. This allows to reduce the number of redundant variables in the subset. Colors in the figure represents the data distribution of the recruitment levels or classes, and the ranking in the right the correlation of each variable with recruitment by means of SUS score. The discriminative power of each variable in relation with these classes can be observed. The first pair of variables (SST and AMO) has the same discriminative power and is redundant. The second pair has higher discriminative power since Anch_SSB allows to discriminate high recruitment level of Hake species and AMO allows to discriminate low recruitment. The third pair is even better because it is able to discriminate all the high recruitments. The final pair its the most discriminative since it also allows to discriminate the medium recruitment.

CFS gives a merit to each feature subset $(X_1, ..., X_z)$, where the correlation of each feature (in the subset) with the class is viewed positively (numerator), whilst correlation between pairs of features (in the subset) is penalised (denominator):

$$Merit(X_1, ..., X_z) = \frac{z \cdot t_{CX}}{\sqrt{z + z(z-1)t_{XX}}}$$

where $\{X_1, ..., X_z\} \subseteq \{X_1, ..., X_n\}$ being z the number of features in the subset, t_{CX} the average class-feature correlation and t_{XX} the average feature-feature correlation of the features included in the subset. Correlation between

two variables is calculated by means of the previously exposed SUS (Hall, 1999).

3.4 Bayesian network classifiers

The kind of practical problems that are addressed in this dissertation need statistical approaches that are characterized by having an explicit underlying probability model, which provides a probability that an instance belongs to each class, rather than simply a classification.

One modelling paradigm based on probability theory and graph theory (Buntine, 1991; Jordan, 1998) is the *Probabilistic graphical models* (PGMs) paradigm (Pearl, 1988; Whittaker, 2009; Thompson, 1992; Lauritzen, 1996; Castillo et al., 1997). PGMs include the particular cases of *Bayesian networks (BNs)* (Jensen and Nielsen, 2001; Neapolitan, 2003; Korb and Nicholson, 2004), that are a paradigm suitable to deal with *uncertainty*, providing an intuitive interface to data. *Bayesian network* models can be used to solve both, *supervised classification* (Friedman et al., 1997; Larrañaga et al., 2005) and clustering problems (Cheeseman and Stutz, 1996).

3.4.1 Bayesian networks

There are many paradigms that have been proposed to the induction of models from labelled data. For example, *classification trees* (Breiman et al., 1984; Quinlan, 1993), where a set of questions are applied hierarchically; the *k-nearest neighbour classifier* searches the database for the most similar k instances in order to classify (Fix and Hodges, 1951); the *neural network classifier* is based on artificial neural networks (McCulloch and Pitts, 1943a), which attempt to mimic the human brain; or, the *support vector machine* (Vapnik, 2000) based on the transformation of the feature space into a higher dimensional space where the different classes can be separated by simple hyperplanes. However, *Bayesian networks* have the advantage of being easier to interpret and extract knowledge than other *supervised classification* paradigms such as *neural networks* (Sebastiani et al., 2005; Correa et al., 2009). However, in domains such as fish recruitment, *probabilistic* classifiers have a useful property for management decision making (Fernandes et al., 2010c). In addition to the forecasting, they also provide the probability of each possible outcome. They have the advantage of being easier to interpret and extract knowledge than other paradigms, due to their graphical representation and their principled probabilistic foundations in domains of high *uncertainty*. This makes them an adequate framework for the necessities of modelling for fisheries management.

Bayesian networks are composed of a graphical representation (structure) where each node corresponds to a variable and arcs (or lack of arcs) represent conditional independence assumptions. In addition to their graphical

representation, they have associated a set of parameters (Jensen and Jensen, 1996; Castillo et al., 1997). Therefore, *Bayesian networks* learning consists of structure and parameter learning. This structure and its parameters can be specified by experts, learned from data or built by combining both (expert knowledge and data learning) (Heckerman et al., 1995).

Plenty of literature about structure learning of graphical models (Calvo, 2008) is available, such as those based on dependency detection algorithms (Chow and Liu, 1968; Schwarz, 1978; Herskovits and Cooper, 1990; Spirtes and Glymour, 1991; Geiger, 1992). However, since the search for the best structure is an NP-hard problem (Chickering et al., 1994; Chickering, 1996), heuristic search methods are often required to obtain structures in a reasonable time. Some of the main approaches are: Greedy search (Buntine, 1991; Cooper and Herskovits, 1992); simulated annealing (Chickering et al., 1995); tabu search (Bouckaert, 1995); genetic algorithms (Holland, 1992; Larrañaga et al., 1996; Etxeberria et al., 1997; Fogel, 2006); estimation of distribution algorithms (Blanco et al., 2003); Markov chain Monte Carlo (Isaacson and Madsen, 1985; Taylor and Karlin, 1998; Myers et al., 1999; Ross, 2007); variable neighbourhood search (De Campos and Puerta, 2001) or; ant colony optimisation (De Campos et al., 2002). Many works about parameter learning have been proposed (Spiegelhalter and Lauritzen, 1990; Heckerman, 1995; Bernardo and Smith, 2001; MacKay, 2003).

In this dissertation, the focus is concentrated on *Bayesian network classifiers* (BNCs), which are a particular kind of *Bayesian network*. BNCs and their advantages are explained in the next section. In addition, the property *Markov blanket* of a *Bayesian network* is also described.

3.4.2 Bayesian network classifiers

Bayesian network classifiers (BNCs) are a subset of the *Bayesian networks* (Larrañaga et al., 2005) which focus their learning on a target (or class) variable (Fig. 3.5). In BNCs, the class variable is the parent of all the features and the number of parents each feature can have is limited. These strong independence assumptions restrict their structure complexity. This has the advantage of allowing efficient and robust learning of structure and parameters. In particular, if there is sparse data available.

As an example, a *naive Bayes* classifier does not allow relationships between features (Minsky, 1961) and *Tree Augmented Naive Bayes* (TAN) only permits one feature parent (Sahami, 1996). The generalization of both BNCs is *k-dBN*, where k is the maximum number of feature parents that are allowed (Friedman et al., 1997). A more flexible representation consists of a 'forest' of tree structures (FAN) rather than a single tree structure (Lucas, 2004).

The *naive Bayes* classifier (Minsky, 1961; Cestnik et al., 1987; Nilsson, 1965; Langley et al., 1992; Duda et al., 2001), one of the most simple *Bayesian network* models for classification (Larrañaga et al., 2005), has been selected

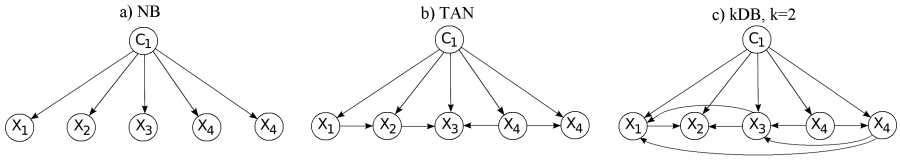


Fig. 3.5. Examples of Bayesian network classifier structures

to be applied in this thesis contribution; this is due to its competitive performance, as it works well in many complex real-world problems (Domingos and Pazzani, 1997; Zhang, 2004). *Naive Bayes* assumes that, given the class variable, all of the factors are independent (Fig. 3.5). This assumption implies that a *naive Bayes classifier* requires the specification of a small number of parameters. This leads to robust models and parameter estimation when sparse training data available (Occam's razor principle; (Domingos, 1999)) as it is common in many marine science problems. Furthermore, it is a computationally-fast model to be learnt (a time complexity of $O(Nz)$), where N is the number of training examples and z is the number of selected factors), which is adequate for wrapper approaches that use the induction algorithm in their search process (Saeys et al., 2007).

3.4.3 Markov blanket

In a *Bayesian network*, the *Markov blanket* (MB) of a node or variable (Pearl, 1988) includes the set of nodes composed by its parents, children and the parents of all of its children. Therefore, the *Markov blanket* of a variable (X) is the smallest set ($MB(X)$) containing all variables carrying information about X, where the provided information can not be increased by adding any other variable (Peña et al., 2007; Pellet and Elisseeff, 2008). This means that the variables of the 'Markov blanket' are the only knowledge base needed to forecast the behaviour of the target variable.

As an example, the *MB* of the recruitment consists of the variables that shield the recruitment from the remaining variables in a *Bayesian Network structure*, i.e. by those variables that, once their value is known, the rest of the variables do not influence the recruitment forecast.

The *MB* can be used as a feature selection method. The *MB* concept can be also used as a post-feature selection process for reducing a large number of variables after a feature selection process (Fernandes et al., 2010a,b).

3.5 Assessing and comparing classification methods

An important issue to consider is the robust estimation of model performance in order to estimate the *uncertainty* of its classification or forecasts and to be able to compare different methodologies or classifiers (Bouckaert and Frank,

2004; Demšar, 2006; García and Herrera, 2008). One common way of assessing the performance of a model consists of comparing its forecasts with available training data. However, a model can easily over-fit the data. This is particularly important if non-parametric and/or optimization methods are used. In addition, there is often other criteria for selecting a particular method, such as interpretability, instead of only performance measures (Alpaydin, 2004).

An additional way of assessing the performance of a model is to check its generalization or forecasting power, i.e. to check how well it behaves forecasting new unseen data. The proper estimation of model performance has been the subject of intensive research (Stone, 1974; Rodríguez et al., 2010).

Therefore, in the following sections several performance measures are presented as well as methods for estimating them robustly. In addition, the issues of comparing methods, meta-learning from the comparison results, and the need to evaluate full pipelines of methods are also addressed.

3.5.1 Classification performance measures

Most performance measures for assessing the quality of classification models are based on the *confusion matrix*. A confusion matrix is a table where the observed counts for each group are presented in the rows, while the model classification is given in the columns (Table 3.3). While a large number of scores are proposed in the literature, in this section, only performance measures used in this dissertation are presented using the example of a boolean class variable.

Table 3.3. Confusion matrix for a boolean class variable

		Predicted class: yes	Predicted class: no
Actual class	yes	true positive (TP)	false negative (FN)
	no	false positive (FP)	true negative (TN)

Accuracy:

Accuracy (Acc), or percent of correctly classified cases is the probability that the model correctly classifies a new instance. *Accuracy* measures model performance considering only the class value with the highest probability (0-1

loss measure), without considering each of the *a posteriori* probability values estimated for each class value.

Accuracy is measured between 0% and 100%, where the highest values indicate the best results. Accuracy is used as the main metric in machine learning because it is a simple way of assessing performance (Pazzani, 1996; Kohavi and John, 1997). The definition of accuracy based on the confusion matrix (Table 3.3) is given by:

$$Acc = \frac{TP + TN}{N}$$

where N is the number of cases or instances.

However, accuracy can be high in datasets where the values of the class variable are not balanced, i.e. one class contains most of the data. Using a model that classifies all the cases within this class (accuracy paradox), the *accuracy* of the classifier is high, but the model is not useful. For this reason, it must be complemented with other performance measures that consider error distribution between all class values such as true positive rate or false positive rate.

True positive rate (TPR):

TPR is the rate of instances that has been correctly classified for a specific class value, i.e. the ratio of positive cases that are correctly classified as positive:

$$TPR = \frac{TP}{TP + FN}$$

It is also known as *recall*, *sensitivity* or *hit rate*.

False positive rate (FPR):

FPR is the rate of instances that are incorrectly classified for a specific class value, i.e. the ratio of negative cases that have been classified as positive:

$$FPR = \frac{FP}{FP + TN}$$

Percent reduction in error (PRE):

The relevance of a performance gain between two methods (or before and after a change in data) can be hard to understand. For example, a 2% accuracy gain of an already highly accurate classifier (e.g. 90%), is not the same as with

a low starting accuracy (e.g. 50%). This can be measured using the Percent Reduction in Error (PRE) (Hagle and Glen, 1992):

$$PRE = 100 \cdot \frac{EB - EA}{EB}$$

where EB is the error in the first method (Error Before) and EA is in the second method (Error After).

Brier score (BS):

Brier score considers the estimated 'a posteriori' probabilities for each possible outcome (Brier, 1950; van der Gaag et al., 2002; Yeung et al., 2005). The *Brier score* for a class variable with r values is given by:

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{r}(p_l^k - y_l^k)^2$$

where N is the number of cases and p_l^k is the forecasted probability for the l^{th} class value for the k^{th} case. The y_l^k value is 1 if l is the observed (correct) value of the class and 0 otherwise. In domains such as recruitment forecasting for fisheries management, the additional information provided by using the *Brier score* is valuable information for the decision-making process (Fernandes et al., 2010c).

The lower the value of BS (between 0 and 2), the better the classifier. However, in this dissertation the BS is divided by 2 in order to have it in the range 0 to 1, which is easier to interpret (Fernandes et al., 2010c). In addition, since it is a measure whose values are difficult to interprete by users, in Table 3.4 a reference based on the author's experience in the recruitment domain is provided (Fernandes et al., 2009b).

Table 3.4. Proposed interpretation of Brier score levels for end-users

Level	Interpretation
> 0.35	Insufficient
≤ 0.35	Acceptable
≤ 0.30	Adequate
≤ 0.20	Superior
< 0.10	Excellent

3.5.2 Performance estimation methods

In order to establish the expected error, the classifier performance has to be assessed; this is accomplished by dividing the user-labelled dataset into two

parts: training and evaluation. Depending upon the selected evaluation technique, this demarcation can be undertaken once or several times, with different data sampling techniques (Rodríguez et al., 2010; Fernandes et al., 2010c).

Hold-out:

One of the simplest approaches consists of leaving part of the data for learning the model, with the remainder used for validation (Larson, 1931). However, the estimated performance can be sensitive to changes in data partition, especially in small datasets (Rodríguez et al., 2010).

K-fold cross-validation:

The sensibility to the data partition can be addressed by data partition into k folds: each fold is left out of the model learning process and used as a test set, repeatedly, k times. The estimated performance is the average of k learned models (Lachenbruch and Mickey, 1968; Stone, 1974; Geisser, 1975).

Repeated k-fold cross-validation:

Such validation processes may be repeated using different partitions of the data, with the results being averaged to ensure replicability. As an example, which has been extensely used during this dissertation, the *10 times-repeated 5-fold cross-validation* (Bouckaert and Frank, 2004) has been coupled with the statistical test *corrected paired t-test* (Nadeau and Bengio, 2003).

Leaving one out cross-validation (LOOCV):

The extreme data partition is to split the data into as many folds as possible, leaving one data instance per fold (Mosteller and Tukey, 1968). This method has the advantage of leaving a large part of the data for learning, but associates a high variance to the reported performance.

Bootstrapping:

High variance can be avoided using re-sampling, with replacement, where the data is re-sampled and the probability of data duplication is considered in the performance estimation (Efron, 1979). However, this approach is computationally expensive.

All approaches have advantages and disadvantages; as such, they must be selected depending upon the data characteristics and validation objectives. However, in general, *stratified 5-fold or 10-fold cross-validation* or its repeated version (*n-times repeated k-fold cross-validation*) stands up as a method that shows a good trade off between robust error estimation and computational

time (Bouckaert and Frank, 2004; Rodríguez et al., 2010). The *stratification* implies that the folds contain approximately the same proportion of class values as the original dataset. Finally, the *n-repeated k-fold cross-validation* consists of repeating n times the cross-validation with different data randomizations. The repeated cross-validation allows avoiding results dependent on data partition, leading to more robust results comparisons using statistical tests; as well as reporting more stable and robust performances by means of averaging the repeats.

Finally, the presented methods are used for error evaluation. However, the same methods can also be used for parameter estimation in order to get a more stable model parameter as used in Schirripa and Colbert (2006) or for more stable feature selection (Francis, 2006; Fernandes et al., 2010c).

3.5.3 The comparison of methods

The choice of which specific learning algorithm should be used is not trivial. In order to accomplish this selection one needs to compare learning algorithms given a dataset in order to assess which algorithm has superior behaviour. However, the issue can be broader, such is the case of comparing multiple algorithms over multiple datasets. In this section, the methodologies used for methods comparison in this dissertation are presented.

3.5.3.1 Comparing multiple classifiers over one dataset

A common method for comparing algorithms is to perform statistical comparisons of the performance measures of trained classifiers on specific datasets (Kotsiantis, 2007). If there are sufficient data, a number of training sets of size N can be sampled, both learning algorithms can be applied to each of them, and the difference in accuracy for each pair of classifiers on a large test set can be estimated.

The next step is to perform a statistical test (e.g. *paired t-test*) to check the null hypothesis that the mean difference between the classifiers is zero. This test can produce two types of errors. Type I error is the probability that the test rejects the null hypothesis incorrectly (i.e. it finds a significant difference although there is none). Type II error is the probability that the null hypothesis is not rejected, when there actually is a difference. The test's Type I error will be close to the chosen significance level.

In practice, however, there is only one dataset of size N and all estimates must be obtained from this single dataset. Different training sets are obtained by subsampling, and the instances not sampled for training are used for testing. Unfortunately this violates the independence assumption necessary for proper significance testing. The consequence of this is that Type I errors exceed the significance level.

Several heuristic versions of the *t-test* have been developed to alleviate this problem (Dietterich, 1998; Nadeau and Bengio, 2003). Ideally, it would

be desirable for the outcome of the test to be independent of the particular partitioning resulting from the randomization process, because this would make it much easier to replicate experimental results published in the literature. However, in practice there is always certain sensitivity to the partitioning used. To measure replicability we need to repeat the same test several times on the same data with different random partitions (usually, 5 or 10 repetitions) and count how often the outcome is the same (Bouckaert, 2003; Bouckaert and Frank, 2004).

In this manuscript the *corrected paired t-test* (Nadeau and Bengio, 2003)) is applied in results from *10-times 5-fold cross-validation*. This test is conservative and can result in higher p-values than other less strict tests (e.g. *paired t-test*).

3.5.3.2 Comparing multiple classifiers over multiple datasets

The comparison of multiple classifiers, or methodologies, can be performed by means of the *revised Friedman plus Shaffer's static post-hoc test*, proposed by García and Herrera (2008) for comparison of multiple methods over multiple datasets. These statistical test results can be represented by means of critical difference diagrams (Deḿsar, 2006), which show the average ranks of the performance of each method across all the domains in a numbered line. If there is not a statistically significant difference between two methods, they are connected in the diagram by a straight line.

As an example, in Figure 6.5, NI and CMindiv methods are connected since they show no significant difference in performance; whereas these methods are unconnected to CMcart , showing a statistically significant difference at the specified levels.

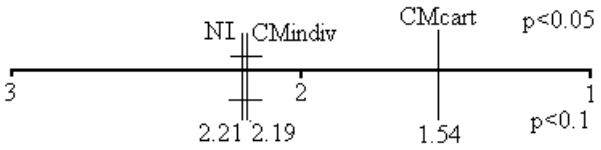


Fig. 3.6. Critical difference diagrams for three missing data imputation methods in synthetic datasets in terms of a performance measure (e.g. accuracy). Methods that do not show a significant difference are connected in the diagram.

3.5.3.3 Meta-learning from performance results

A key question when dealing with classification is not whether a learning algorithm is superior to others, but under which conditions a particular method

can significantly outperform others on a given application problem. Meta-learning is moving in this direction, trying to find functions that map datasets to algorithm performance (Kalousis et al., 2004).

Meta-learning uses a set of attributes, called meta-attributes, to represent the characteristics of learning tasks, and searches for the correlations between these attributes and the performance of learning algorithms. As an example, characteristics of learning tasks are: the number of instances, the proportion of categorical attributes, the proportion of missing values, the entropy of classes and others (Brazdil et al., 2003; Fernandes et al., 2010b).

3.5.4 Pipeline performance evaluation

As already introduced, often pipelines of supervised methods are applied to data. In order to avoid model over-fitting and provide an honest validation, the entire pipeline has to be included in the validation scheme (Reunanen, 2003; Statnikov et al., 2005). This means that the data partition, in folds, is performed before the application of the first step of the pipeline. Therefore, not only the classification model, but also all the pipeline is validated. In order to validate a proposed pipeline in this dissertation, a 10 times repeated 5-fold cross-validation (10x5cv) schema has been selected as recommended in (Bouckaert and Frank, 2004; Rodríguez et al., 2010).

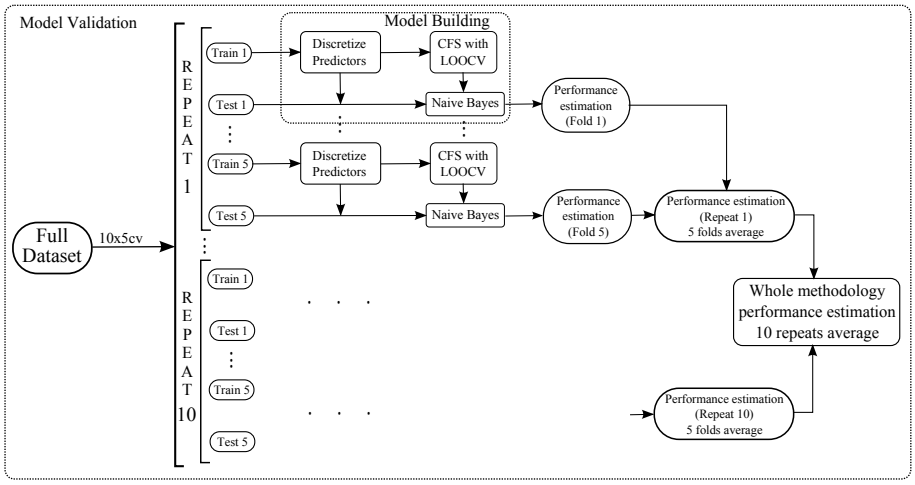


Fig. 3.7. Example of a validation scheme for a pipeline that contains only filter methods

As an example, in Figure 5.4 a pipeline is validated, which is composed of a feature discretization step, a feature selection step and a *naive Bayes* classifier.

However, this process can be more complex if there is a wrapper step, i.e. a step that uses a model to perform its task, with the necessity of performance evaluation in the process. In this case, this pipeline contains a loop of validation for the wrapper method (Fig. 3.7), which is part of the model building. If such a pipeline is validated, the data partition (inner loop) for the wrapper process is performed inside an outer loop of validation for the pipeline evaluation. Therefore, the inner loop only has access to the subset of data provided by the outer loop. Such a pipeline is used in Fernandes et al. (2010c).

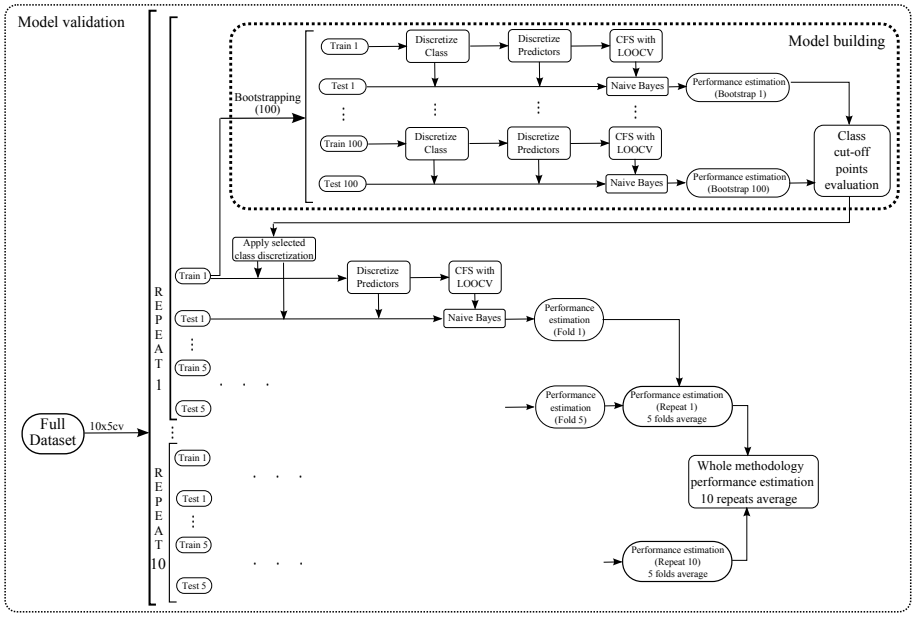


Fig. 3.8. Example of a validation scheme for a pipeline that contains a wrapper step

3.6 Multiple class variables classification (multi-dimensional)

In this section, the supevised classification framework for multi-dimensional (Mul-D) domain is introduced and the case of the *naive Bayes* classifier, generalized to the simultaneous prediction of multiple class variables is also described. In addition, the currently available performance measures are also introduced.

3.6.1 Multi-dimensional supervised classification

The objective of multi-dimensional *supervised classification* (Mul-D) (van der Gaag and de Waal, 2006; de Waal and van der Gaag, 2007; Rodríguez and Lozano, 2008) consists of building a classifier from training data S, with N cases (Table 3.5), in order to forecast the value of a vector of m class variables $\mathbf{C} = (C_1, \cdots, C_m)$, instead of just one single class variable, given the vector of feature variables $\mathbf{X} = (X_1, \cdots, X_n)$ of an unseen unlabelled case $\boldsymbol{x} = (x_1, \cdots, x_n)$. As an example, in fish recruitment, $(C_1, \cdots, C_m)$ could represent different fish species recruitment to be forecasted and $(X_1, \cdots, X_n)$ represents the factor set (climatic, biological and others) or features. This approach profits from the class-variables that can be related between them. Therefore, simultaneous forecasting of all class variables can accomplish better results than by separate forecasting. In addition, it would be often desirable to model them together for the interpretation of experts, instead of modelling each species in separate models (Uni-D).

Table 3.5. Data matrix of a multi-dimensional *supervised classification* problem.

	X_1	...	X_n	C_1	...	C_m
Instance 1	x_1^1	...	x_n^1	c_1^1	...	c_m^1
Instance 2	x_1^2	...	x_n^2	c_1^2	...	c_m^2
Instance 3	x_1^3	...	x_n^3	c_1^3	...	c_m^3
...		...			...	
Instance N	x_1^N	...	x_n^N	c_1^N	...	c_m^N

3.6.2 Multi-dimensional naive Bayes classifier

The choice of *naive Bayes* is mainly motivated by the fact that *naive Bayes* for one class variable problems, or uni-dimensional (Uni-D) classification, has outperformed other more complex paradigms within the fish recruitment forecasting domain (Fernandes et al., 2010c), where data is usually scarce.

In the previously introduced *uni-dimensional naive Bayes classifier* (UDnB), the joint distribution $p(\mathbf{x}, c)$ can be expressed as:

$$p(x_1, \cdots, x_n, c) = p(c) \prod_{i=1}^{n} p(x_i|c)$$

In a *multi-dimensional naive Bayes classifier* (MDnB), the class variables are the parents of all the features, the classes have no parents and the features have no other feature as a parent (Fig. 3.9). In the case of a MDnB, the joint probability distribution $p(\mathbf{x}, \mathbf{c})$ is given by:

$$p(X_1, \cdots, X_n, C_1, \cdots, C_m) = \prod_{j=1}^{m} p(c_j) \prod_{i=1}^{n} p(x_i|c_1, ..., c_m)$$

In order to classify a case $\mathbf{x} = (x_1, \cdots, x_n)$ into a vector of class variables $\mathbf{c} = (c_1, \cdots, c_m)$, the *joint classification rule* (Rodríguez and Lozano, 2008, 2010), which returns the most probable value for each class variable simultaneously, has been selected:

$$(\hat{c}_1, \cdots, \hat{c}_m) = argmax_{c_1, \cdots, c_m} \{p(c_1, \cdots, c_m|x_1, \cdots, x_n)\}$$

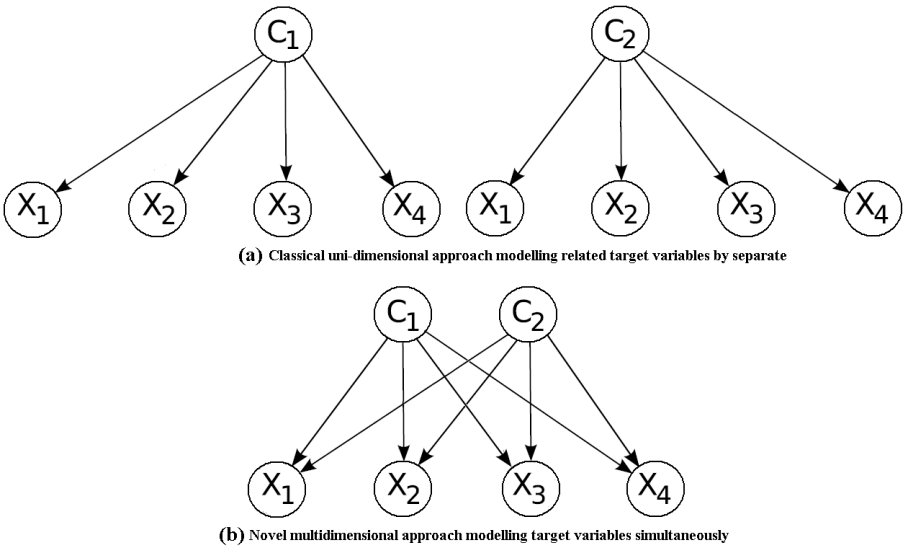


Fig. 3.9. Examples of uni-dimensional (Uni-D) and multi-dimensional classifiers (Mul-D).

3.6.3 Multi-dimensional performance measures

There are few performance measures in the literature specific for this multi-dimensional approach. To the best of our knowledge there is the so-called *joint accuracy* (Rodríguez and Lozano, 2010) proposed in (van der Gaag and de Waal, 2006). This is where a case is classified correctly, if all the class variables are labelled correctly simultaneously. The author proposal and adaptation of performance measures for the multi-dimensional approach are presented in Part II.

Finally, the uni-dimensional classification approach is assumed throughout the dissertation. Otherwise it is explicitly specified that a multi-dimensional domain or framework is being described.

Part II

Advances in supervised classification for fisheries research

This second part of the dissertation shows the author contributions to the application of *supervised classification* methods to marine science problems related with fisheries management and how these problems have been formulated considering the particular characteristics of each domain:

- Chapter 4 proposes a wrapper method for helping experts in deciding the number of classes or taxa in zooplankton classification.
- Chapter 5 presents a *supervised classification* application to single fish species recruitment forecasting with the following objectives; a) forecasts with its *uncertainty* associated; b) forecasts and scenarios easy to interpret; c) search for recruitment and factors boundaries that can be interpreted; d) high factors stability; e) Error balanced through all recruitment levels and; f) robust error estimation.
- In Chapter 6, the simultaneously multiple fish species recruitment forecasting by means of the multi-dimensional classification approach is presented. In this chapter, a set of 'state-of-the-art' uni-dimensional pre-processing methods, within the categories of missing data imputation, feature discretization and feature subset selection, are adapted to be used for multi-dimensional classifiers. Those proposed methods are tested with synthetic datasets and the real domain of fish recruitment.

In order to ensure reproducibility of methods and results (Buckheit and Donoho, 1995; Barnes, 2010), the used datasets and Java implementations of the methods used in this dissertation are available from the ISG group webpage (*www.sc.ehu.es/ccwbayes/members/jafernandes/*). The computers used in the experiments consists of a simple dual core processor 2.0 GHz with 4GB of RAM memory or lower hardware configuration.

4

Optimizing the number of classes in zooplankton classification

4.1 Introduction

Zooplankton biomass and abundance estimation by size spectrum or taxa (e.g. Fig. 4.1) is carried out routinely in marine research. Zooplankton plays a key role in the transference of primary production to fish and it is important to understand marine ecosystems (Irigoien et al., 2002, 2004).

However, the analysis of plankton samples is costly in experts time. Therefore, machine-learning techniques for the identification of plankton, combined with automated or semi-automated image analysis processes for feature extraction (e.g. Table 4.1), have been proposed to assist in sample analysis.

However, a difficulty in automated plankton recognition and classification systems is the selection of the number of classes or taxa. The end-user wants the maximum number of taxonomical detail as well as the minimum recognition error. Therefore, a methodology that allows the end-user to find a good trade-off between classification performance and taxa detail is needed.

A method that combines human knowledge with machine-learning techniques is proposed (Fernandes et al., 2009c), in order to allow the end-user to have help in the labelling of zooplankton images.

The aim is to maximize both the performance of the classifier and the number of classes while maintaining the meaningful information for the end-user. In the proposed method, a machine-learning method provides the results of performance and the number of classes, whereas the end-user provides the ecologically meaningful information and the initial maximum number of classes.

4.2 Zooplankton datasets

The method has been applied to three different example datasets (Table 4.2).

A high resolution (2400dpi) public dataset (Tulear_04 dataset) available at the ZooImage webpage (*www.sciviews.org/zooimage*) has been selected in order to permit reproducibility (Fig. 4.2).

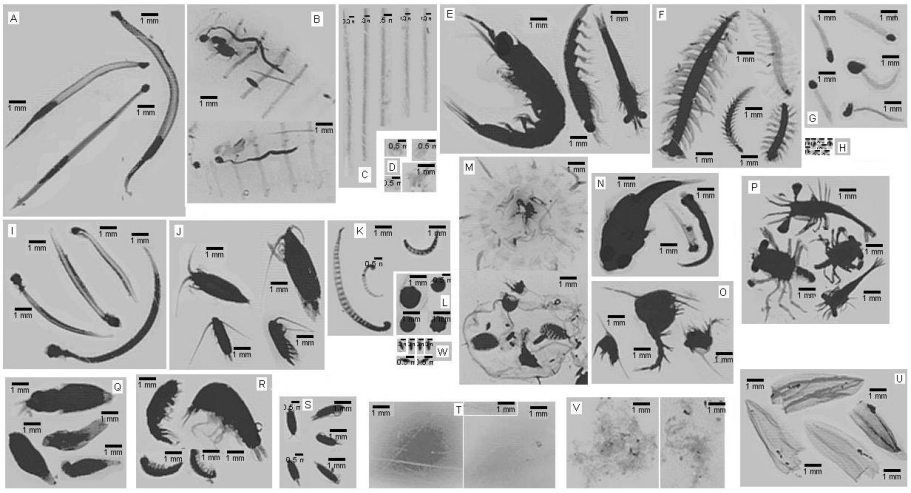


Fig. 4.1. Images representative of each taxa presented in a dataset from the Bioman oceanographic surveys (98-06; at 600dpi). A: Chaetognaths. B: Doliolids. C: Artifacts from scanning process. D: Small marine snow. E: Euphausiids. F: Polychaeta. G: Appendicularia. H: Small zooplankton. I: Fish larvae. J: Large copepoda. K: Polychaeta larvae. L: Round zooplankton. M: Gelatinous. N: Fish. O: Decapod larvae type I. P: Decapod larvae type I. Q: Salps. R: Crustacean others. S: Medium copepoda. T: Stained jelly. U: Siphonophora. V: Marine snow. W: Small copepoda. X: Multiple copepoda.

In addition, two datasets from bioman oceanographic surveys are used. Bioman_98-06 dataset has been established with zooplankton samples scanned at 600 dpi; Bioman_07 dataset with 2400 dpi images. Both datasets have been built from samples obtained in the Bay of Biscay preserved in 4% borax buffered formalin, then stained with eosin. Eosin staining avoids the imaging of inorganic debris in the image analysis step through image filters. This staining process can reduce the number of artifact particles between a 50% and a 75% (unpublished data for Bachiller (2008)). All the datasets were analyzed using ZooImage framework. However, any methodology, framework or tools for data acquisition and processing preferred by the expert can be used.

The variables considered were those routinely extracted by ZooImage, together with a limited number of environmental variables for Bioman_98-06 dataset (Table 4.1). Expert taxonomists labelled the images of each dataset with the aim of achieving the maximum number of classes possible. In Tulear_04 dataset, 1639 individuals were classified into 37 classes. For the Bioman_98-06 dataset, 17803 individuals were classified into 24 classes. For Bioman_07 dataset, 6724 were classified into 30 classes (Table 4.2, Fig. 4.2 and 4.1).

The datasets are used for illustration purposes; the method can be applied to datasets obtained with any other methodology.

Table 4.1. Individual features: Morphological and image measurements extracted by ZooImage, using the image analysis software, ImageJ. Environmental features collected during the survey can be added.

Feature	Description
	ZooImage (ImageJ) features
ECD	Equivalent circular diameter
Area	Surface area
Mean	Mean of the gray scale of the pixels
Skew	The third-order moment, about the mean of the gray scale
Kurt	The fourth-order moment, about the mean of the gray scale
Std. dev.	Standard deviation of the gray scale of the pixels
Mode	Mode of the gray scale of the pixels
Median	Median of the gray scale of the pixels
Min.	Minimum of the gray scale of the pixels
Max.	Maximum of the gray scale of the pixels
IntDen	Sum of the gray values of the pixels
XM	Coordinate horizontal of the gray scale center of the pixels
YM	Coordinate vertical of the gray scale center of the pixels
Perim.	Perimeter
Width	Width of the rectangle, containing the individual
Height	Height of the rectangle, containing the individual
Major	Longest axis of the ellipsis, containing the individual
Minor	Smallest axis of the ellipsis, containing the individual
Circ.	Circularity
Feret	Diameter of longest distance between the two points of the individual
	Environmental features
Temperature	Surface temperature
Salinity	Salinity of the sample
Depth	Depth of the sample
Latitude	Latitude of the sample
Longitude	Longitude of the sample

4.3 Method for optimizing the number of classes and classification performance

The methodology proposed consists of three steps, as outlined below.

1. The end-user distributes the extracted images of individuals into all the groups, which can be visually identified (i.e. labeling). A classifier is trained with this dataset and the corresponding estimated performance is used as a starting point.
2. All possible mergers of *two classes* into a *single class* are evaluated. For each pair of classes, a new dataset is constructed, in which the two classes

Table 4.2. Number of individuals per class in the three used datasets before any merger.

First dataset:	Tulear_2004	Second dataset:	Bioman_1998-2006	Third dataset:	Bioman_2007
# of individuals	Classes	# of individuals	Classes	# of individuals	Classes
27	Bubble	467	Artifact	110	Artifact
50	Scratch	482	Small marine snow	97	Small marine snow
50	Shadow	1136	Marine snow	97	Medium marine snow
50	Debris	2228	Small copepoda	49	Large marine snow
50	Diatom	2063	Medium copepoda	198	Small copepoda
50	Fiber	2361	Large copepoda	207	Copepoda multiple
50	Marine snow	871	Multiple copepoda	2288	Calanoida
50	Other phytoplankton	1838	Euphasiacea	1189	Cyclopoida oncaea
50	Calanoida dorsal I	208	Decapoda larvae I	110	Cyclopoida corycaeaus
49	Calanoida dorsal II	122	Decapoda larvae II	548	Cyclopoida oithona sp
50	Calanoida dorsal III	279	Polychaeta	86	Cyclopoida oithona nana
50	Calanoida lateral	12	Polychaeta larvae I	168	Harpaticoida microsetella
50	Eucalanidae	31	Amphipoda	208	Harpaticoida euterpina
39	Temoridae	209	Appendicularia	174	Appendicularia
50	Oithonidae	1123	Chaetognatha	115	Chaetognatha
39	Miraciidae	107	Doliolida	12	Euphausiacea
50	Corycaeidae	202	Siphonophorae	32	Decapoda larvae
50	Oncaeidae	57	Hydroidomedusae	244	Cladocera
50	Poicilo lateral	160	Stained jelly (rests)	28	Nematoda
8	Sapphirinidae	17	Cephelopoda larvae	250	Doliolid
50	Annelida	48	Pisces	20	Siphonophora
22	Cirripeda	200	Pisces larvae	84	Hydroidomedusae
50	Cladocera	3043	Zooplankton small	142	Bivalvia larvae
26	Decapoda miscellaneous	539	Round zooplankton	18	Gastropoda
50	Decapoda zoea dorsal			58	Pteropoda
50	Decapoda zoea lateral			32	Polychaeta
50	Malacotraca bulky			52	Copepoda egg I
50	Elongated malacostraca			46	Copepoda egg II
21	Malacostraca larvae			16	Fish Egg
22	Cnidaria			16	Diatom
37	Appendicularia				
50	Chaetognatha				
50	Elongated egg				
49	Round egg				
50	Protista				
50	Gastropoda				
50	Pisces				
1639	37	17803	24	6694	30

are merged into a unique class, whereas the remainder are left unchanged. A classifier is constructed from this new dataset and its performance is evaluated. The possible mergers are ranked, based on their estimated performance (e.g. accuracy, but other can be used). Optionally, the confusion matrix (CM) can be used to reduce the number of mergers to be evaluated using the classes with more misclassified individuals counts above a certain threshold (e.g. mean of non-zero misclassified in the CM; Table 4.3). This option significantly reduces computation time. Step 2 is automatically performed by a computer program (Fig. 4.3), which outputs a ranking (Table 4.5) with all possible mergers of two classes and their associated statistics (see below). The method implementation uses Weka API algorithms to perform these steps (Witten and Frank, 2005).

Fig. 4.2. Images representative of each taxa presented in the original Tulear_04 dataset (scanned at 2400dpi). Bubble (A), Scratch (B), Shadow (C), Debris (D), Diatom (E), Fiber (F), Marine snow (G), Other phytoplankton (H), Calanoida dorsal I (I), Calanoida dorsal II (J), Calanoida dorsal III (K), Calanoida lateral (L), Eucalanidae (M), Temoridae (N), Oithonidae (O), Miraciidae (P), Corycaeidae (Q), Oncaeidae (R), Poicilo lateral (S), Sapphirinidae (T), Annelida (U), Cirripeda (V), Cladocera (W), Decapoda miscelaneus (X), Decapoda zoea dorsal (Y), Decapoda zoea lateral (Z), Malacostracea bulky (AA), Elongated malacostraca (AB), Malacostraca larvae (AC), Cnidaria (AD), Appendicularia (AE), Chaetognatha (AF), Elongated egg (AG), Round egg (AH), Protista (AI), Gastropoda (AJ), Pisces (AK).

In order to ensure reproducibility (Buckheit and Donoho, 1995), a Java implementation of the method is available from the ISG group webpage (*www.sc.ehu.es/ccwbayes/members/jafernandes/*).

3. The end-user evaluates the ranking of mergers and decides which specific mergers to accept considering not only the performance that can be achieved, but also the ecological value of the new merged class and the objective of the research study. A new classifier, with end-user selected mergers, is trained and evaluated. This new classifier can be compared

with those learned in the first step and in previous iterations. The end-user can perform steps 2 and 3, repeatedly.

```
While User does not end mergers evaluation
  Build classifier before mergers
  Evaluate classifier
  Calculate metrics (accuracy, ...)
  Save classifier metrics in mergers ranking
  For all i ∈ CLASS 1, CLASS 2, ..., CLASS n − 1
    For all j ∈ CLASS i + 1, ..., CLASS n
      If ((CM) and (CLASS i and CLASS j in CM list)) or (not CM) then
        Reset dataset to original without mergers
        Merge CLASS i and CLASS j in dataset
        Build classifier with merged dataset
        Evaluate classifier
        Calculate metrics (accuracy, ...)
        Save classifier metrics in mergers ranking
      End if
    End for
  End for
  Perform user selected mergers
End while
```

Fig. 4.3. Pseudocode used to describe the method. Pseudocode is not language-programming dependent; and it omits programming details that are not relevant to specify the method. CM represents if the use of confusion matrix has been selected or not.

The method proposed relates to optimizing the number of classes (class selection) and the classification performance. Therefore, it can be applied to data from any source and classified with different methods as long as they are classified into different classes that can be grouped without losing all the information (e.g. grouping different taxonomic levels). The method could be run 'manually' but the expert would be confronted with hundreds of mergers to explore without previous knowledge of the potential accuracy gain. Any merger does not lead to an accuracy gain; in fact, there is a high rate of mergers that decrease performance (Table 4.4). Automation and ranking of the results leave only a limited number of mergers, with the highest accuracy, for the end-user to analyze; as opposed to the end-user manual 'trial and error' exploration without previous knowledge of the potential performance gain.

The method is independent of any specific machine learning paradigm for classification or evaluation as well as any specific performance metric. The end-user can select different classification paradigms and performance metrics taking into account the specific requirements of the study being undertaken (e.g. taxonomic groups, compared with ecological impact). In our examples, a *Tree Augmented Naive Bayes classifier* (TAN) was used for classification (Friedman et al., 1997). The TAN classifier is used for a faster mergers evaluation. It shows a good performance record, lying close to Random Forest. Random Forest proved to be a good classification algorithm for zooplankton (Grosjean et al., 2004).

Table 4.3. Confusion matrix of the classifier before mergers evaluation for Tulear_04 dataset. The main diagonal in the centre represents the correctly classified individuals. The rest of the cells are the misclassified individuals. Columns show the classifier classification and rows show the user labelled class present in the dataset. The selected mergers by the end-user are displayed in gray.

User↓ | Classifier→	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
Oncaeidae (A)	**29**	0	3	4	8	0	0	0	0	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
Round egg (B)	0	**38**	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	1	1	0	2	0	0	1	1
Calanoida lateral (C)	6	0	**12**	1	0	5	0	0	0	0	8	1	0	1	1	2	3	0	0	0	0	0	0	0	0	3	2	2	2	0	0	0	0	1	0	0	0
Corycaeidae (D)	1	0	3	**18**	0	12	1	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	8	0	4	0	0	0	0	0	0	0	0	0
Miraciidae (E)	0	0	0	0	**38**	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Poicilo lateral (F)	9	0	0	0	13	0	**18**	0	0	0	0	2	0	0	1	0	0	1	0	0	0	0	0	0	0	3	0	1	2	0	0	0	0	0	0	0	0
Eucalanidae (G)	0	0	0	1	0	0	**40**	0	0	3	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Scratch (H)	0	0	0	0	0	0	0	**47**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Debris (I)	0	3	0	0	0	0	0	0	**24**	0	0	0	0	0	0	0	0	0	5	0	0	0	0	1	0	6	2	0	0	0	1	3	4	0	0	0	1
Calanoida dorsal III (J)	0	0	0	0	1	0	7	0	0	**34**	0	5	2	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Calanoida dorsal I (K)	2	0	5	1	0	1	3	0	0	0	**24**	5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	5	0	0
Calanoida dorsal II (L)	3	0	1	0	0	0	1	0	0	6	3	**22**	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	1	9	0	0
Fiber (M)	0	0	0	0	1	0	1	1	0	0	0	0	**40**	0	0	0	0	0	0	0	4	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0
Decapoda miscellaneous (N)	0	0	0	0	0	0	0	0	0	1	0	2	0	**11**	0	0	1	0	0	0	0	0	0	0	0	0	7	4	0	0	1	3	0	0	0	0	0
Appendicularia (O)	0	0	1	0	0	0	0	0	0	0	0	2	1	0	**21**	0	0	0	2	0	5	2	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0
Decapoda zoea lateral (P)	2	0	1	0	0	1	0	0	0	0	1	0	0	1	0	**27**	5	0	0	0	0	0	0	0	0	2	0	4	4	0	2	0	0	0	0	0	0
Decapoda zoea dorsal (Q)	3	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	**30**	0	0	0	0	0	0	0	0	1	0	5	1	0	1	0	0	0	1	0	0
Bubble (R)	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	**16**	0	0	0	1	2	1	3	0	0	0	0	1	0	0	0	0	0	0	0
Annelida (S)	0	0	2	0	1	0	2	0	3	1	2	1	0	1	0	3	0	0	**24**	0	0	0	0	0	0	1	0	1	3	0	1	0	0	1	0	3	0
Sapphirinidae (T)	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	1	0	0	3	0	0	0	1	0	0
Oithonidae (U)	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2	0	0	0	0	0	0	**40**	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Cirripeda (V)	0	0	2	0	0	0	0	0	0	0	1	0	1	0	2	0	1	0	0	0	0	3	**5**	0	0	2	0	0	1	0	0	2	0	0	2	0	0
Shadow (W)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	**47**	0	0	0	0	0	1	0	0	0	0	0	0	0
Diatom (X)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**43**	5	0	0	0	0	0	2	0	0	0	0	0
Protista (Y)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	**47**	0	0	0	0	0	0	0	0	0	0	0	0
Pisces (Z)	4	3	1	1	0	2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**37**	0	0	0	0	1	0	0	0	0	0	0
Elongated egg (AA)	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**48**	0	0	0	0	0	0	0	0	0	0
Malacostraca bulky (AB)	0	0	1	1	0	0	0	0	0	0	0	0	0	0	5	0	0	0	1	0	0	0	0	0	0	0	0	**38**	2	0	0	0	0	0	1	1	0
Marine snow (AC)	1	0	4	0	0	2	0	0	2	0	3	1	0	5	1	2	0	0	2	0	0	0	0	0	0	1	0	6	**15**	0	0	0	0	1	1	3	0
Cnidaria (AD)	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	**14**	0	2	0	1	0	1	2
Cladocera (AE)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	**41**	0	0	0	0	0	0
Other phytoplankton (AF)	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	3	0	0	0	0	0	0	1	0	1	1	0	1	0	2	1	**29**	0	0	0	0	3
Gastropoda (AF)	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	**44**	0	0	0	0
Malacostraca larvae (AH)	4	0	6	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	3	0	0	2	0	0	0	**1**	0	0	0
Temoridae (AI)	2	0	1	0	0	0	3	0	0	0	8	17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**7**	0	0
Elongated malacostraca (AJ)	0	0	2	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	**44**	0
Chaetognatha (AF)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	**48**

Table 4.4. For each iteration, several statistics are presented after performing the end-user selected mergers. 'Before' represents accuracies before performing any merge. The number of evaluated mergers is represented by '#Mergers'. Accuracy is the overall accuracy after performing selected mergers. The 'p-values' are the result of performing an statistical comparison (*corrected paired t-test*) between datasets with different number of classes. E.g. the 'p-value original' refers to the comprarison with the original dataset before any merger; whereas the 'p-value previous' refers to the comparison with the resulting dataset of the previous iteration. The same with PRE score that is provided, for both, in relation with the previous iteration dataset and in relation with the original. 'Mergers↓' is the rate of mergers that instead of improving accuracy reduces it. 'CPU-time' is the computer time to evaluate the mergers. 'CM' corresponds to statistics when using the confusion matrix to reduce the mergers to be evaluated.

Merger evaluation		Tulear_04	Bioman_98-06	Bioman_07
Before	Accuracy (%)	64.7	85.7	82
After fist iteration	Accuracy (%)	68.3	87.3	82.1
	P-value original	0.585	0.078	0.976
	PRE original (%)	10.2	4.7	0.6
	#Mergers selected	4	5	4
	#Mergers evaluated	666	276	435
	Mergers↓ (%)	78.3	21.4	91
	CPU-time	3:01:39	0:32:34	1:30:47
	CPU-time CM	0:17:37	0:16:07	0:17:31
	#Mergers evaluated CM	58	29	33
After second iteration	Accuracy (%)	70.9	88.8	-
	P-value previous	0.542	0.7	-
	P-value original	0.395	0.006	-
	PRE previous (%)	8.2	4.6	-
	PRE original (%)	17.6	9	-
	#Mergers selected	4	1	-
	#Mergers evaluated	628	190	-
	Mergers↓ (%)	74.7	63.7	-
	CPU-time	1:57:40	0:17:45	-
After third iteration	Accuracy (%)	73	-	-
	P-value previous	0.514	-	-
	P-value original	0.179	-	-
	PRE previous (%)	7.2	-	-
	PRE original (%)	23.5	-	-
	#Mergers selected	2	-	-
	Mergers↓ (%)	69	-	-
	CPU-time	1:41:29	-	-
After fourth iteration	Accuracy (%)	73.9	-	-
	P-value previous	0.426	-	-
	P-value original	0.699	-	-
	PRE previous (%)	3.3	-	-
	PRE original (%)	26.1	-	-
	#Mergers selected	1	-	-
	Mergers↓ (%)	16.9	-	-
	CPU-time	1:01:32	-	-
After fifth iteration	Accuracy (%)	74	-	-
	P-value previous	0.679	-	-
	P-value original	0.398	-	-
	PRE previous (%)	0.4	-	-
	PRE original (%)	26.3	-	-
	#Mergers selected	1	-	-
	Mergers↓ (%)	20.4	-	-
	CPU-time	0:40:37	-	-

Table 4.5. Ranking of mergers with the highest accuracies for Tulear_04 dataset, at the first iteration. In each row, the accuracy, the PRE, the classes to be merged and the user-decision are given.

Top 10 mergers iteration 1	User decision
66.5%, PRE: 5.1%, Oncaeidae with Calanoida lateral	X: One is Poecilostomatoida, the other Calanoida
66.3%, PRE: 4.5%, Corycaeidae with Poicilo lateral	$\surd$: Both are Poecilostomatoida
66.2%, PRE: 4.2%, Scratch with Temoridae	X: Scratch is an artifact and Temoridae is not
66.0%, PRE: 3.7%, Oncaeidae with Poicilo lateral	$\surd$: Both are Poecilostomatoida
65.9%, PRE: 3.4%, Eucalanidae with Calanoida dorsal III	$\surd$: Both are Calanoida
65.8%, PRE: 3.1%, Decapoda miscellaneous with Decapoda zoea lateral	$\surd$: Both are Decapoda
65.8%, PRE: 3.1%, Decapoda Miscellaneous with Malacostraca larvae	X: One is Decapoda and the other Malacostracea
65.7%, PRE: 2.8%, Decapoda Zoea lateral with Pisces	X: One is Decapoda the other Pisces
65.7%, PRE: 2.8%, Decapoda Zoea dorsal with Gastropoda	X: One is Decapoda the other is Gastropoda
65.6%, PRE: 2.5%, Decapoda Miscellaneous with Malacostraca bulky	X: One is Decapoda and the other Malacostracea

In terms of validation, 5-fold cross-validation has been considered sufficient to suggest the mergers (Kohavi, 1995; Rodríguez et al., 2010).In order to assess the classifier performance, several measures are used in addition to accuracy; percent reduction in error (PRE), true positive per class (TP), false positive per class (FP). In addition, the statistical *corrected paired t-test* (Nadeau and Bengio, 2003) to asses if differences in performance are statistically significant is used. While the *corrected paired t-test* shows interesting properties with respect to its non-corrected version (e.g. overlap between training folds is taken into account), it shows a more conservative behaviour (higher p-values). Accuracy, overall correctly classified instances, is used as the main metric because it is a simple way of assessing performance (Pazzani, 1996; Kohavi and John, 1997). However, the end-user can define other metrics depending on the study objectives. Finally, the relevance of a performance gain can be hard to understand. For example, a 2% accuracy gain on an already high accuracy scenario (e.g. 90%) is not the same as with a low accuracy (e.g. 50%). This relevance can be measured using the PRE score (Hagle and Glen, 1992). $PRE = (100 \cdot (EB - EA)/EB)$, where EB is error before mergers and EA error after mergers. TP is the proportion of individuals that have been correctly classified as belonging to a class. Similarly, FP is the proportion of individuals that not being of a certain class are incorrectly classified as being part of it.

4.4 Application examples

The evaluation of the new classifiers with class mergers is shown in Tables 4.4 and 4.6. In Tulear_04 dataset, 64.7% accuracy was obtained with the initial 37 classes. Out of 666 possible two-class mergers considered, 145 (21.7%) showed an improvement in accuracy. The list of the mergers, which resulted in the highest improvement in accuracy, was evaluated by a end-user who accepted five mergers (Fig. 4.4). Several iterations were performed until there were no further mergers accepted by the user. After the third iteration, there was an 8.3% accuracy gain, a PRE of 23.5%. However, the improvement in accuracy may not be significant enough (p<0.20). In Bioman_98-06 dataset, there is a 3.1% accuracy gain after two iterations, a PRE of 9%, with differences that are statistically significant (p<0.05). In Bioman_07 dataset, there is a small accuracy gain of only 0.1% and the new classifier is not significantly different from the previous one (p>0.05).

4.5 Discussion

The proposed method consists of a semi-automated process of possible mergers, which can balance both objectives, i.e. the maximization of the number of

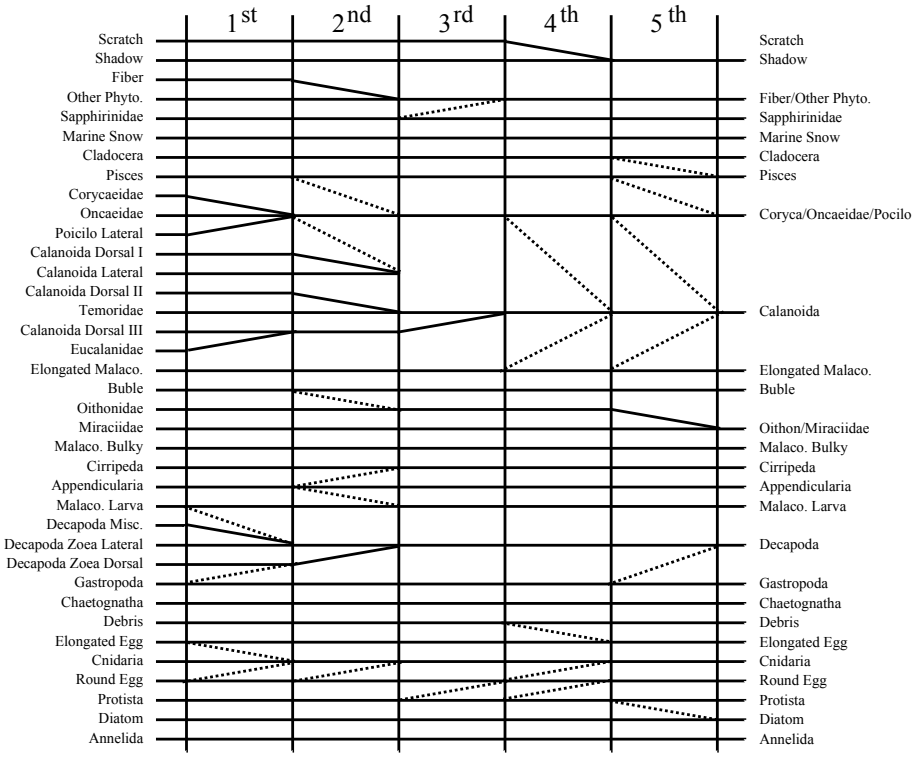


Fig. 4.4. Graphical representation of accepted class mergers in Tulear_04 dataset by the end-user for each iteration. Accepted mergers are represented by straight lines; whereas, in doted lines, some machine proposed mergers rejected by the end-user are shown.

classes and the performance. The exhaustive study of all possible class combinations is computationally unfeasible, i.e. not only the merge of classes in pairs, but also in triplets, quads and others. As an example, the number of possible combinations for Tulear_04 dataset (37 classes) is $3.74409 \cdot 10^{43}$. The total number of class mergers to be evaluated (two-classes mergers + three-classes mergers + four-classes mergers $+ \cdots +$ $(n-1)$-classes mergers) can be calculated by means of Stirling numbers of second kind (Abramowitz and Stegun, 1964):

$$X = \left(\sum_{k=0}^{n} S(n,k) \right) - 2$$

(excluding 'not performing any merger' and 'merging in a unique class'). In this expression, X is the number of possible combinations, n is the number of classes to consider for possible mergers and $S(n,k)$ is broken down as

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^n$$

This number of combinations could be reduced if only two-class mergers were evaluated in each iteration and the process performed repeatedly:

$$\sum_{k=n}^{3}\binom{k}{2} = \binom{n}{2} + \binom{n-1}{2} + \binom{n-2}{2} + \cdots + \binom{3}{2}$$

As an example, the number of possible two-class mergers, evaluated in the first iteration in Tulear_04 dataset, is 666. If only one merger is performed, the next iteration evaluates 630 mergers. However, if the end-user decides to perform four mergers, this results in 33 taxa or classes in the next merging iteration, with 528 possible mergers to evaluate. In spite of this reduction in the number of evaluations, it remains a computationally expensive task (hours of CPU-time for Tulear_04 dataset first iteration, Table 4.4) that can be reduced using the confusion matrix to find a good set of merger candidates instead of trying all two-class mergers (<20 min of CPU-time, Table 4.4).

Occasionally, more than one merger per iteration could lead to a lower accuracy. However, this has never been observed during these experiments and several mergers per iteration are selected by the user to speed up the process.

The proposed method presents a number of benefits: (i) the end-user has a framework within which to accomplish a 'trade-off' between the number of classes and performance; (ii) the absence of monotonicity between the number of classes and accuracy can result in improved performance for more detailed datasets. (iii) the user can avoid testing mergers that actually decrease performance.

The particular objectives of each end-user's study have an impact on the decision of accepting or rejecting mergers. However, the end-user faces the question of whether the accuracy gains obtained after merging classes are relevant or not. The proposed metrics (accuracy, PRE, TP, FP and the p-value) should help in taking such decisions and to evaluate classifiers effectiveness. The following example using Tulear_04 dataset illustrates a possible use of these metrics (Table): the accuracy gain is not significant after the third iteration, so the end-user could make use of the classifier obtained at that step. However, the TP rate of Oithonidae and Miraciidae improves with the classifier obtained after the fifth iteration (Table 4.6). If these classes were important for the end-users research, the decision would be to select the classifier obtained after the fifth iteration. Most of merged classes in all datasets present significant improvements in TP and FP with little variations in the rest of the classes.

Table 4.6. Classifier overall accuracy (correctly classified), TP rates and FP rates, per class in each classifier (generated after 'end-user' selected mergers, in each iteration) for Tulear_04 dataset. For a given class value, TP rate is the percentage of individuals classified in a class by the classifier, which belong to that class in the training set. FP is the percentage of individuals classified as belonging to a class when they are not. TP and FP experiment low variation in classes not being merged and high improvement in most of the merged classes.

	Before mergers		After iteration 1		After iteration 2		After iteration 3		After iteration 4		After iteration 5	
Accuracy	0.647		0.683		0.697		0.73		0.739		0.74	
PRE	-		0.102		0.142		0.235		0.261		0.263	
Classes	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Bubble	0.593	0	0.556	0.002	0.63	0.001	0.63	0.003	0.593	0.003	0.481	0
Scratch	0.94	0.001	0.96	0.002	0.96	0.001	0.96	0.002	0.94	0.001	0.95	0.001
Shadow	0.94	0.004	0.86	0.002	0.9	0.001	0.84	0.001				
Debris	0.48	0.009	0.54	0.006	0.54	0.009	0.56	0.006	0.52	0.006	0.52	0.009
Diatom	0.86	0.003	0.86	0.003	0.88	0.004	0.86	0.003	0.88	0.004	0.88	0.004
Fiber	0.8	0.006	0.82	0.008	0.71	0.022	0.73	0.025	0.75	0.022	0.77	0.025
Other phytoplankton	0.58	0.01	0.52	0.01								
Marine snow	0.3	0.02	0.28	0.014	0.3	0.013	0.24	0.013	0.28	0.013	0.26	0.013
Calanoida dorsal I	0.48	0.022	0.36	0.022	0.41	0.038	0.813	0.069	0.816	0.064	0.806	0.066
Calanoida lateral	0.24	0.021	0.18	0.012								
Calanoida dorsal II	0.449	0.023	0.408	0.021	0.648	0.026						
Temoridae	0.179	0.013	0.282	0.014								
Calanoida dorsal III	0.68	0.009	0.86	0.024	0.87	0.021						
Eucalanidae	0.8	0.015										
Oithonidae	0.8	0.012	0.72	0.011	0.6	0.01	0.48	0.008	0.62	0.006	0.73	0.006
Miraciidae	0.974	0.002	0.923	0.002	0.897	0.001	0.897	0.003	0.846	0.001		
Corycaeidae	0.36	0.014	0.847	0.052	0.813	0.044	0.8	0.056	0.827	0.058	0.807	0.054
Oncaeidae	0.58	0.023										
Poicilo lateral	0.36	0.021										
Sapphirinidae	0	0	0	0	0	0.001	0	0	0	0	0	0
Annelida	0.48	0.009	0.54	0.006	0.5	0.005	0.5	0.006	0.5	0.006	0.5	0.006
Cirripeda	0.227	0.004	0.318	0.005	0.273	0.006	0.273	0.003	0.227	0.003	0.318	0.004
Cladocera	0.82	0.004	0.86	0.006	0.82	0.004	0.84	0.004	0.84	0.004	0.84	0.004
Decapoda miscellaneous	0.423	0.01	0.539	0.025	0.651	0.034	0.651	0.03	0.667	0.03	0.659	0.029
Decapoda zoea lateral	0.54	0.012										
Decapoda zoea dorsal	0.6	0.011	0.56	0.009								
Malacostraca bulky	0.76	0.02	0.76	0.016	0.74	0.014	0.7	0.014	0.76	0.014	0.78	0.015
Elongated malacostraca	0.88	0.008	0.9	0.008	0.92	0.006	0.9	0.004	0.9	0.005	0.9	0.004
Malacostraca larvae	0.048	0.002	0.095	0.001	0	0.001	0.095	0.001	0	0.001	0.095	0.001
Cnidaria	0.636	0.003	0.591	0.005	0.636	0.004	0.591	0.006	0.545	0.005	0.591	0.006
Appendicularia	0.568	0.007	0.514	0.009	0.514	0.006	0.514	0.009	0.514	0.009	0.514	0.006
Chaetognatha	0.96	0.004	0.92	0.004	0.94	0.004	0.94	0.004	0.96	0.004	0.92	0.003
Elongated egg	0.96	0.003	0.98	0.004	0.98	0.003	0.98	0.004	0.96	0.003	0.96	0.003
Round egg	0.776	0.004	0.755	0.004	0.755	0.004	0.755	0.003	0.776	0.003	0.776	0.003
Gastropoda	0.88	0.004	0.84	0.004	0.86	0.004	0.88	0.003	0.86	0.004	0.88	0.004
Protista	0.94	0.007	0.94	0.005	0.94	0.006	0.94	0.005	0.92	0.004	0.96	0.005
Pisces	0.74	0.021	0.7	0.015	0.6	0.014	0.56	0.013	0.52	0.011	0.56	0.012

4.6 Conclusions and suggestions for future work

The proposed method allows to reduce the end-users *uncertainty* with respect to training-set elaboration, by providing guidance to balance the number of classes and the classification performance. The end-user can initially separate all the identifiable groups, check the mergers decision in terms of automatic classification and then evaluate the proposed changes according to performance (accuracy, PRE, TP, FP or significance of the improvements)

and the research objectives. Lastly, the method is independent of any specific machine learning technique, but simple techniques are selected and a code implementation is provided. Future work will focus on the automation of mergers exploration and on the unbalanced nature of zooplankton datasets.

The expose research has been published and leaded to the following research contributions in refereed journals and international forums:

- (2009) **Optimizing the number of classes in automated zooplankton classification**. *Fernandes J.A.*, Irigoien X., Boyra G., Lozano J.A. and Inza I. Journal of Plankton Research 31(1): 19-29.
- (2009) **Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment**. Irigoien X., *Fernandes J.A.*, Grosjean P., Denis K., Albaina A. and Santos M. Journal of Plankton Research 31(1): 1-17. Featured article.
- (2009) **Changes in plankton size structure and composition, during the generation of a phytoplankton bloom, in the central Cantabrian sea**. Zarauz L., Irigoien I. and Fernandes J.A. Journal of Plankton Research. 31(2): 193-207.
- (2008) **Modelling the influence of abiotic and biotic factors on plankton distribution in the Bay of Biscay, during three consecutive years (2004-06)**. Zarauz L., Irigoien X. and Fernandes J.A. Journal of Plankton Research 30(8): 857-872.
- (2008) **Relevance of otolith features for anchovy age classification**. Ascoreca A., *Fernandes J.A.*, Cotano U., Uriarte A. and Irigoien X. Eleventh International Symposium on Oceanography of the Bay of Biscay. Donostia - San Sebastian (Spain). April 2-4th.
- (2010) **A comparison between digital camera and scanner as imaging devices for semi-automated zooplanktton classification using microscope classification as control**. Bachiller E., *Fernandes J.A.* and Irigoien X. Aquatic Sciences: Global Changes from the Center to the Edge. International Joint Meeting with ASLO & NABS. Santa Fe, NM, USA. June 6-11th.
- (2009) **FlowCAM/PhytoImage intercalibration exercise**. Denis K., Tunin-Ley A., Fernandes J. A., Maurer S., Parent J.-Y., Belin C., Irigoien X. and Grosjean Ph. Third SCOR WG130 meeting. Baton Rouge (LA), USA. May 13-26th.
- (2010) **Small-scale vertical distribution of zooplankton in the Catalan Sea: Relationships with physical characteristics**. Alcaraz M., Saiz E., Lebourges-Dhaussy A., Graña R., Cotano U., *Fernandes J.A.*, Isari S., Zamora S., Mouriño B. and Irigoien X. Rapp. Comm. Int. Mer Medit. Vol. 39 - pp. 84.
- Perfil (CTM2006-12344-C02-02) project from the Spanish Ministry of Education and by the Department of Agriculture, Fisheries and Food of the Basque Country Government.
- (2009) Perfil oceanographic surveys at Mediterranean Sea. Flowcam. June.

- (2008) Perfil oceanographic surveys at Cantabrian Sea. Flowcam. July.
- (2008) Bioman oceanographic surveys at Cantabrian Sea. Pairovet and Flowcam. May.
- (2007) Bioman oceanographic surveys at Cantabrian Sea. Cufes and Flowcam. May.

In addition, as an expert on this area the author has provided training and counseling to international institutions:

- (2010) France: Delphine Bonnet and Juliette Balavoine (CNRS-Universit Montpellier II).
- (2010) Taiwan: Elise Marquis (National Taiwan University).
- (2009) Spain: Sonia Romero Romero (Universidad de Cdiz).
- (2009) Spain: Cristina Garcia Muñoz (CSIC).
- (2008-2009) Spain: Mara Lidia Nieves (Universidad de las Palmas).
- (2008-2009) Spain: Pablo Len (Universidad de Mlaga).
- (2008-2009) Spain: Enric Saiz (CMIMA), CSIC.
- (2008) France: Serge A. Poulet (INRS).
- (2008) Greece: Nikolaos Nikolioudakis (University of Crete).
- (2008) South Africa: Fabienne Cazassus (MA-RE Institute, UCT).
- (2007) Portugal: Mara Manuel Angelico (IPIMAR).

Finally, the authors work in this area has appeared in Basque Country television through a collaborator in the science divulgate program *Kresala*.

5

Advances in fish recruitment forecasting by means of supervised classification

Improving the ability to forecast fish recruitment is a key element in fisheries management. However, the interactions between population dynamics and different environmental factors are complex and often non-linear, making it difficult to produce robust forecast (Uriarte et al., 2002).

Machine learning techniques (in particular, supervised classification methods) have been proposed as useful tools to overcome such difficulties (Dreyfus-León and Chen, 2007; Dreyfus-León and Schweigert, 2008). However, several methodological issues have been rised by fisheries experts mainly due to the sparse data available which lead to unstable results (Allain et al., 2001; Uusitalo, 2007).

In this study, a methodology is proposed (Fernandes et al., 2010c) to build a robust classifier for fish recruitment forecast with sparse and noisy data.

The methodology consists of 4 steps: 1) a semi-automated recruitment discretization method; 2) a supervised discretization of factors; 3) a multivariate and non-redundant factors selection; and 4) learning a probabilistic classifier.

In terms of fisheries management, the estimated classifier performance has important consequences and, to be useful, the manager needs to know the risk that is being taken when using this estimation. In addition, probabilistic classifiers such as *naive Bayes* have the advantage that, in addition to the forecast, the estimated probability of each recruitment level is provided.

Anchovy (Engraulis encrasicolus) and hake (Merluccius merluccius) recruitments are used as application examples in this study. 'Two-intervals' recruitment discretization accomplishes 70% accuracy rate and Brier scores of around 0.10, for both anchovy and hake recruitment. In comparison, 'three-intervals' recruitment discretization accomplishes 50% accuracy rate; and Brier scores of around 0.25 for anchovy and 0.30 for hake recruitment. These statistics are the result of validating not only the classifier, but also the previous steps, as a whole methodology (Reunanen, 2003; Statnikov et al., 2005).

The principal objective of this study is to propose a machine learning based framework, to perform a probabilistic forecast of recruitment with techniques and results that are robust and useful for management decisions (providing

stable and replicable performance estimations, as well as forecast *uncertainty* estimates).

The proposed methodology is a pipeline of state-of-the-art machine-learning methodologies, addressing several critical steps: recruitment and forecast discretization; factors selection; performance estimation and final model learning.

There are two main potential end-users of this methodology: 1) the scientist who could use the methodology as a data mining tool to find out variables that might affect recruitment and investigate related hypotheses and; 2) the fisheries manager that might have a tool to evaluate risks for the fishery.

In order to be able to use probabilistic classification models, the target variable has to be discretized (Torgo and Gama, 1997; Frank et al., 2000; Revoredo and Zaverucha, 2004), i.e. the regression problem has to be transformed into a classification problem. As an example, in marine science fish recruitment values are often discretized using equal width (Dreyfus-León and Chen, 2007), or equal frequency.

Nevertheless, these methods produce artificial boundaries, which do not have any biological or management meaning. Therefore, the discretization is undertaken often on the basis of fisheries experts suggestions.

However, sometimes insufficient information about the effects on the model performance is available for setting these boundaries (Uusitalo, 2007). As recruitment boundary decisions affect dramatically the final model and the consequent results (Uusitalo, 2007), in the proposed methodology a recruitment discretization method is included (class discretization in data mining literature).

The proposed approach considers the number of intervals, the domain significance of the cut-off points and the balance of the number of instances within each interval. All of these are critical issues identified in Uusitalo (2007), which must be addressed in order to ensure robustness and usability of the final model.

5.1 Methods

5.1.1 Application examples

The method has been applied to two species of commercial interest in the Bay of Biscay: anchovy (Engraulis encrasicolus) and hake (Merluccius merluccius).

These are two cases where stock-recruitment relationships are poor factors of recruitment (Fig. 5.1.1), where research on climate-recruitment relationships has been undertaken.

Anchovy recruitment and climate have been the subject of intensive studies (Motos et al., 1996; Bellier et al., 2007; Allain et al., 2007; Borja et al., 2008; Planque and Buffaz, 2008). Hake recruitment relationship with climate and spawning stock biomass (SSB), which has been associated with temperature

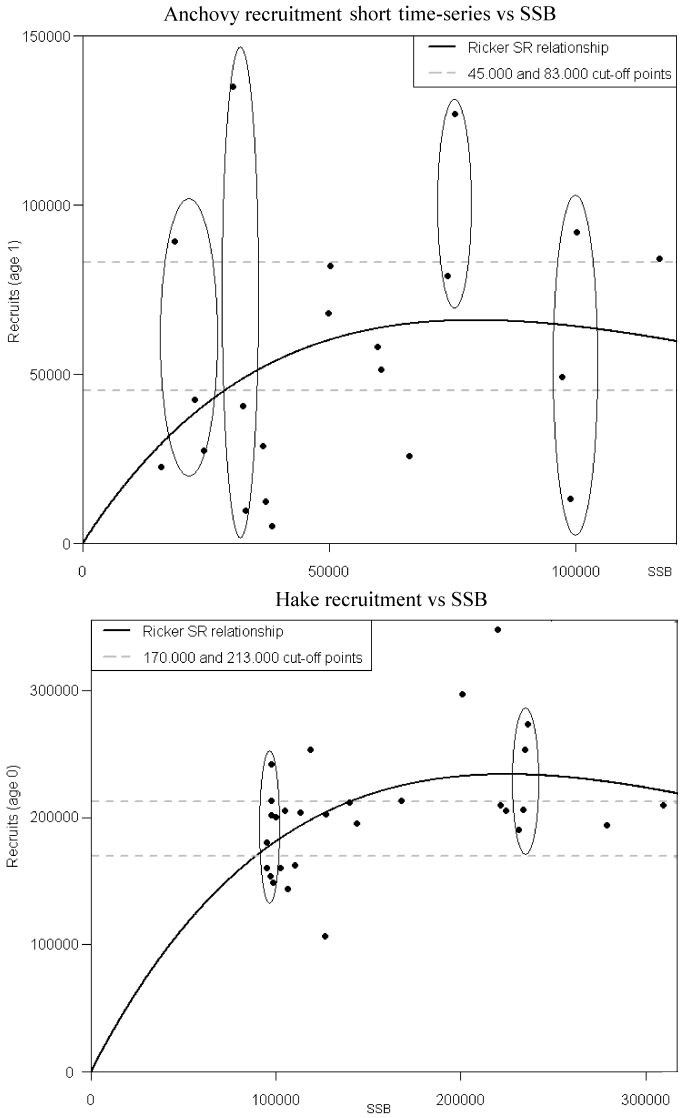


Fig. 5.1. Relationships between spawning stock biomass and recruitment, for anchovy in the Bay of Biscay and for the northern hake stock. The solid black line indicates the adjustment of a Ricker model. The dotted lines shows the cut-off points suggested by the model, for recruitment discretization. The elliptic selections are examples of why SSB is inefficient in discriminating recruitment.

in other studies (McFarlane et al., 2000; Bartolino et al., 2008), is the subject of ongoing research (Meiners, 2007).

5.1.2 Data sources

The target variables to forecast are the Anchovy Recruitment Index long time-series (ARI; Borja et al. (1996)), Anchovy Recruitment (AR; Uriarte et al. (2008b)) and the Hake Recruitment (HR; Uriarte et al. (2008a)). The datasets are the result of working groups of the International Council for the Exploration of the Sea (ICES).

In the case of anchovy, ARI is a recruitment index time-series (1967-2005; 39 years, Fig. 5.2) established from the percentage of age 1 in the landings, but where there is not spawning stock biomass estimations for the period 1967-1987. AR in the Bay of Biscay is available only from 1987 to the present (21 years) from a two-stage biomass dynamic model (Ibaibarriaga et al., 2008).

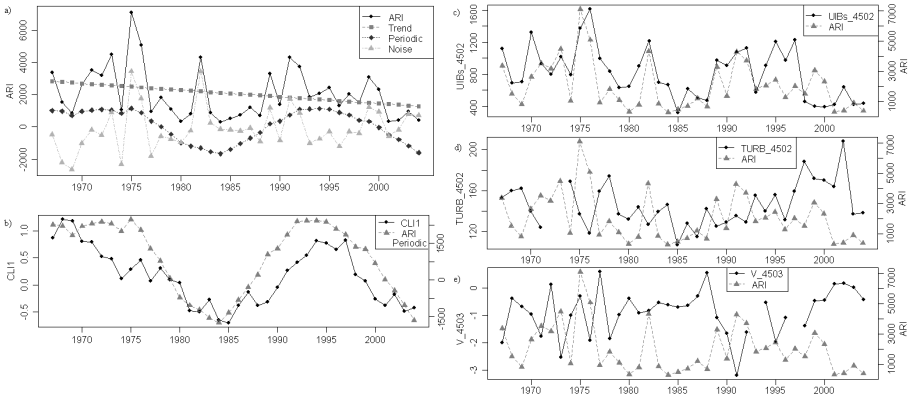


Fig. 5.2. Anchovy recruitment index (ARI) time-series for the period 1967–2005, compared to its selected factors time-series: (a) ARI different time-series components: trend, periodicity and noise; (b) the 6-year running mean of the CLI1 and ARI periodic component comparison; (c) UIBs_ 4502 and ARI; (d) TURB_4502 and ARI; (e) V_4503 and ARI time-series.

The AR is somewhat short time-series for data mining studies. However, the proposed methodology has been also applied to this short recruitment time-series, to evaluate the role of the spawning stock biomass (SSB). Finally, The HR time-series consists of 29 years of data (1978-2006).

In addition to SSB, a set of environmental variables made available by the experts are used (Table 5.1).

The main dataset of environmental variables used in this study has been obtained from the 2007 Workshop on 'Long-term Variability in SW Europe' (ICES, 2007). The compiled variables consist of global climatic and physical oceanographic indices, proposed by experts researching these species.

Climatic indices for the Atlantic region were represented by the key modes of large-scale atmospheric circulation over the northern hemisphere (Barnston

Table 5.1. Sets of variables considered for each specie recruitment forecast. Many variables considered for hake recruitment have not been considered for anchovy, since there is no anchovy presence in the areas where those variables are measured.

Factors or features	Anchovy	Hake
1. Global climatic indices:		
NAO, EA, WP, EP_NP, PNA, EA_WR, SCA, TNH, POL, PT, NAO_DM, NAO_m, At_global and At_NH.	√	√
2. Global climatic indices from PCA analysis:		
CLI1, CLI2 and CLI3.	√	√
3. Physical oceanographic indices:		
AMO, SSTP, RFG, SSTSS, POLE, UIs_4311, Uim_4311, TPEA, SST_4503, SST_4311, TAIR_4311, U_4503, V_4503, U_4311, V_4311, NWPw, NWPs, SOFWE, SUFWE, HSFWE, SONFWE, UILm_4502, UIBm_4502, UIBs_4502, TURB_4502, HF_4503, LHF_4503, ZMF_4503 and MMF_4503.	√	√
4. Other climatic indices:		
TempAnomGlobal (hadsst2), TempAnomNH (hadsst2), Natlantic.average, AMO (unsmoothed), WinterNAO.NOAA, SpringEA.NOAA and Central England temperature.	√	√
5. Solar activity:		
Sunspot and AA_index.	√	√
6. Regional temperature indices:		
TempAnom: A (4045N 5-0W), B (40-45N 10-5W), C (4550N 50W), D (45-50N 10-5W), E (45-50N 15–10W), F (50-55N 0-5E), G (5055N 5-0W), H (50-55N 10-5W), I (50-55N 10-5W), J (55-60N 5-10E), K (55-60N 0-5E), L (55-60N 5-0W), M (55-60N 10-5W), N (55-60N 15-10W), O (60-65N 0-5E) and P (60-65N 0-5E).		√
7. Local wind indices:		
E-W (46.5N 4.5W), N-S (46.5N 4.5W), E-W (48.5N 9.5W), N-S (48.5N 9.5W), E-W (50.5N 7.5W), N-S(50.5N 7.5W), E-W (53.5N 12.5W), N-S (53.5N 12.5W), E-W (57.5N 8.5W), N-S (57.5N 8.5W), E-W (61.5N 4.5W), N-S (61.5N 4.5W), E-W (58.5N 1.5E) and N-S (58.5N 1.5W).		√

and Livezey, 1987). Six indices were selected for analysis: North Atlantic Oscillation (NAO); East Atlantic pattern (EA); East Atlantic/Western Russia pattern (EA/WR); Scandinavia pattern (SCA); Tropical/Northern Hemisphere pattern (TNH); and Polar/Eurasia pattern (POL).

These indices, covering the period 1950-2006, were obtained from the US National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (*www.cpc.ncep.noaa.gov*). De-trended series of climate variables were analysed using principal component analysis, to extract the main patterns of variability, following Varimax rotation. Subsequently, climatic variability was represented by the first three principal components, accounting for up to 63.8% of the total variance. However, the first two components (CLI1 and CLI2) accounted for only 22.3 and 22.1% of the variance, respectively; this indicates that none of the climatic indices were prevalent (see details in Bode et al. (2006)).

In addition to the afore mentioned dataset, other climatic indices have been added: winter (Dec-March) mean NAO index, since it shows its max-

imum fluctuations over this season; spring (March-July) mean EA pattern, coinciding with the anchovy spawning period in the Bay of Biscay (Borja et al., 1996, 1998, 2008); and 'global' mean temperature values for the whole North Atlantic and the northern hemisphere.

Two solar indices have been also considered (*www.ngdc.noaa.gov*): annual number of Sunspots; and Sun geomagnetic activity (AA_index).

In the case of hake, some additional variables have been considered to represent the whole distribution area: Sea Surface Temperature (SST) anomalies, obtained from the hadSST2 dataset (Rayner et al., 2006), on a 5x5 grid-box basis over the northern hake distribution area; and vectorial data of wind, in particular, east-west (u) and north-south (v) geostrophic wind components obtained from NOAA (Fleet Numerical Oceanographic 8, *www.pfeg.noaa.gov*), from which vectorial wind data can be estimated (Table 5.1).

5.1.3 Model-building

The proposed methodology consists of performing supervised factors discretization, followed by a supervised factors selection (in a leaving one out cross-validation scheme) and finally learning a *naive Bayes classifier*.

The approach can be applied to a dataset where the values of the recruitment have been discretized by the end-user (Fig. 5.3), or the recruitment discretization (class discretization) can be part of the proposed model-building process (Fig. 5.4), in a bootstrap scheme.

Finally, the whole methodology (pipeline of supervised classification methods) is validated by means of 10 times-repeated 5-fold cross-validation (10x5cv; Fig. 5.3 and 5.4).

Recruitment semi-automated discretization (class discretization):

A semi-automated recruitment discretization methodology is proposed, in order to establish optimal cut-off point sets for recruitment. In the proposed semi-automated discretization method, a ranking of cut-off point sets is compiled with their associated estimated performance measures.

This ranking is presented to the fisheries expert, who has to select the final cut-off point set to be adopted (Table 5.3), who selectes the cut-off point sets that are useful for management or knowledge extraction.

The performance of each cut-off point set is estimated using 100 resampling sets in a 0.632*bootstrap* schema (Efron, 1979) over the full model-building proposed schema: supervised factors discretization, factors selection and a *naive Bayes classifier* (model-building; Fig. 5.3). In the present study, two criteria for the cut-off point set selection are investigated: fixing the objectives in the maximization of the mean of true positive rate (*max_mean_tp*); or in the maximization the accuracy (*max_accuracy*).

Finally, the method can evaluate all recruitment cut-off point set combinations, or the cut-off point set candidates can be restricted. I.e., the expert

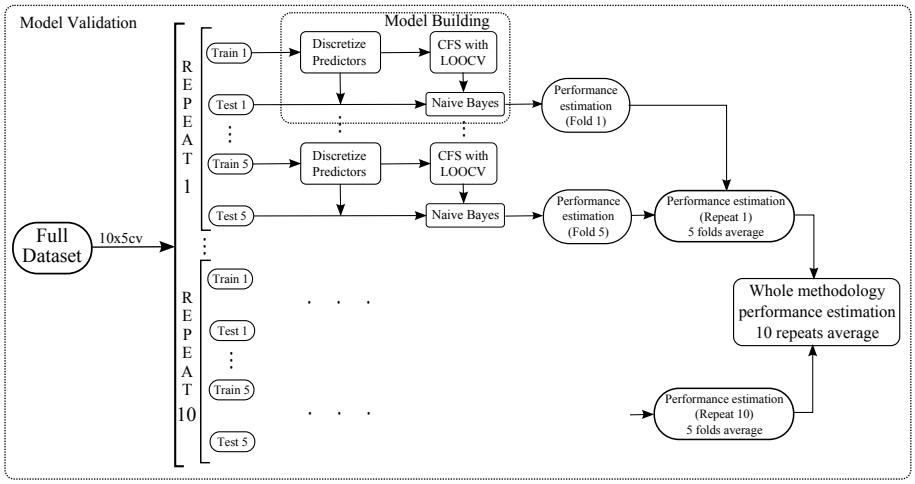


Fig. 5.3. Validation scheme for an end-user defined recruitment discretization. The 'model-building' consists of: the supervised discretization of factors; a multivariate factors selection, 'in a leaving one out' validation scheme (CFS with LOOCV); and learning a *naive Bayes classifier*. A 10x5cv validation scheme is used, in order to estimate the performance of this methodology. For a honest validation every step in 'model-building' only uses training data and the validation is performed using unseen test data.

Table 5.2. List of recruitment cut-off point sets proposed by the discretization algorithm presented to the end-user for anchovy recruitment. (Note: the table shows only a part of the whole list of cut-off point sets proposed by the algorithm). The ranking can be composed of hundreds or thousands of possible cut-off point sets. The end-user selected cut-off point sets are shaded and the maximum performance scores are in bold. 'St.' is the abbreviature for stability of the CFS feature selection method, which counts the number of times that the subset of features is selected.

Anchovy recruitment index discretization in 3 bin											
Wrapper *max_mean_TP* or *max_accuracy* (100 bootstrapps)											
CutPoint1	CutPoint2	#Inst1	#Inst2	#Inst3	TP1	TP2	TP3	TPmean	Acc.	CFS st.	CFS selected subset
1200	3250	10.4	10.2	6.4	0.81	0.7	**0.81**	**0.77**	77.4	**34**	POL; CLI1; V_4503; UIBs_4502
1500	3250	12.3	8.2	6.4	**0.87**	0.73	0.7	0.76	**80.2**	29	POL; CLI1; TURB_4502
1050	2550	9.4	9.1	8.6	0.83	**0.75**	0.71	0.76	76.6	20	PT; CLI1; NAO_DM; PEA
1100	3250	9.9	10.7	6.4	0.79	0.72	0.75	0.76	75.7	34	POL; CLI1; UIBs_4502
1100	3150	9.9	10.1	7.1	0.86	0.74	0.64	0.75	76.1	19	UIs_4311; V_4503; UIBs_4502; AA_Index

can limit the cut-off point sets evaluated to those that have a minimum number of instances per interval by setting a threshold. Yang and Webb (2009) propose the use of a threshold equal to the square root of the total number of instances, in order to establish a minimum number of instances in each factor interval,. Although suggested in the factors discretization literature, it is extended here to the proposed recruitment discretization method. As a result, the problem of the imbalance in the number of instances in each recruitment

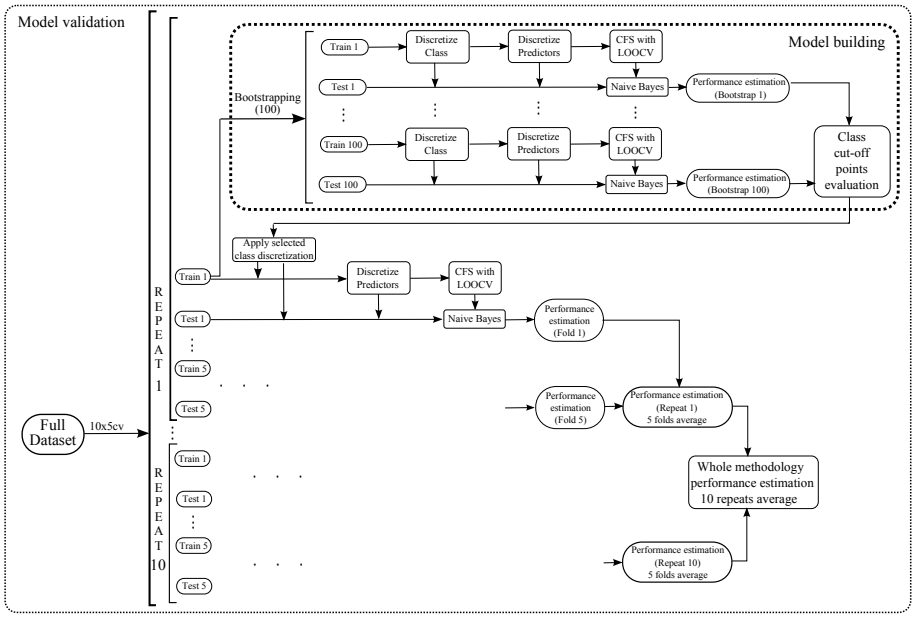


Fig. 5.4. Example of a validation scheme for a pipeline that contains a wrapper step

Table 5.3. List of recruitment cut-off point sets proposed by the discretization algorithm presented to the end-user for anchovy recruitment. (Note: the table shows only a part of the whole list of cut-off point sets proposed by the algorithm). The ranking can be composed of hundreds or thousands of possible cut-off point sets. The end-user selected cut-off point sets are shaded and the maximum performance scores are in bold. 'St.' is the abbreviature for stability of the CFS feature selection method, which counts the number of times that the subset of features is selected.

Anchovy recruitment index discretization in 3 bin											
Wrapper *max_mean_TP* or *max_accuracy* (100 bootstrapps)											
CutPoint1	CutPoint2	#Inst1	#Inst2	#Inst3	TP1	TP2	TP3	TPmean	Acc.	CFS st.	CFS selected subset
1200	3250	10.4	10.2	6.4	0.81	0.7	**0.81**	**0.77**	77.4	**34**	POL; CLI1; V_4503; UIBs_4502
1500	3250	12.3	8.2	6.4	**0.87**	0.73	0.7	0.76	**80.2**	29	POL; CLI1; TURB_4502
1050	2550	9.4	9.1	8.6	0.83	**0.75**	0.71	0.76	76.6	20	PT; CLI1; NAO_DM; PEA
1100	3250	9.9	10.7	6.4	0.79	0.72	0.75	0.76	75.7	34	POL; CLI1; UIBs_4502
1100	3150	9.9	10.1	7.1	0.86	0.74	0.64	0.75	76.1	19	UIs_4311; V_4503; UIBs_4502; AA_Index

interval is avoided, whilst CPU-time is reduced.

Factors supervised discretization:

The variables (factors) have been discretized using the state-of-the-art Fayyad and Irani's MDL Multi Interval Discretization (MID) method (Fayyad and Irani, 1993). This approach is a supervised method that searches for cut-off point sets, minimising the recruitment entropy given each factor (condi-

tional entropy). Entropy is a measure of *uncertainty* (Shannon and Weaver, 1963), whilst the conditional entropy quantifies the discrimination power of a factor, in relation to recruitment. The lower the conditional entropy the better the discrimination power of a factor (lower *uncertainty*).

In addition, the Fayyad and Irani's method has a Minimum Description Length (MDL) penalisation criterion, to avoid selecting a large number of intervals. The method can discard variables, setting a unique interval for all its values, when there is 'no-discretization' that reduces significantly the *uncertainty* of the recruitment value. These variables are removed for subsequent analysis.

Factors selection:

The multivariate Correlation-based Feature Selection (CFS) method has been adopted as a prior step to classifier learning (Hall and Smith, 1997; Hall, 2000). The CFS formulation is based upon the assumption that a good subset of factors (features in the data mining literature) is the one where each of its factors is highly correlated with the recruitment; and at the same time, the factors have low correlation between them. CFS searches for a subset of factors that is relevant for the recruitment, where the factors are non-redundant between them or this redundancy is minimal. CFS gives a merit to each factor set, where the correlation of each factor (in the set) with the recruitment is viewed positively (numerator), whilst correlation between factors (in the subset) is penalised (denominator):

$$Merit(X_1, ..., X_z) = \frac{z \cdot t_{CX}}{\sqrt{z + z(z-1)t_{XX}}}$$

where k is the number of factors in the subset, t_{CX} is the average recruitment-factor correlation and t_{XX} is the average factor-factor correlation. Correlation between two variables is calculated by means of the classical Symmetrical Uncertainty Score (SUS), bounded between 0 and 1 (Hall, 1999).

In addition to the results of CFS, a univariate ranking between each factor and recruitment using the non-parametric Symmetrical Uncertainty Score (Hall, 1999) has been calculated. The univariate ranking is not part of the proposed methodology, but can be of interest for the expert; this is performed in order to examine variables that are highly correlated with the recruitment, that could have not been selected in the described multivariate factors selection.

Finally, the most repeated subset of variables selected by CFS in a 'leaving one out cross-validation scheme' (LOOCV) is more robust than performing CFS directly on all the data (Francis, 2006). In this way, the most selected factor subset in the LOOCV scheme is considered as the most stable subset, ensuring a more robust set of variables. This stability is needed in this kind of research where data is costly to collect and selected variables have important biological meaning (Kalousis et al., 2005, 2007; Kuncheva, 2007).

However, the fisheries expert has the final decision on which set of variables is selected for building the final model, from the suggested ranking provided by the cross-validated CFS (Table 5.4).

Supervised classification model:

Bayesian networks have the advantage of being easier to interpret and extract knowledge than other supervised classification models such as 'Neural networks' (Correa et al., 2009), due to their graphical representation and their principled probabilistic foundations in domains of high *uncertainty* (Sebastiani et al., 2005). *Naive Bayes* (Duda and Hart, 1973; Langley et al., 1992), one of the simplest *Bayesian network* model for classification (Larrañaga et al., 2005), has been selected; this is due to its competitive performance, as it works well in many complex real-world problems (Domingos and Pazzani, 1997; Zhang, 2004). *Naive Bayes* assumes that, given the recruitment or class variable, all of the factors are independent. This assumption implies that a *naive Bayes classifier* requires the specification of a small number of parameters. Further, it is a computationally-fast model to be learnt (a time complexity of $O(nk)$, where n is the number of training examples and k is the number of selected factors). This is adequate for wrapper approaches that use the induction algorithm in their search process (Saeys et al., 2007).

Another advantage of the *naive Bayes classifier* (and probabilistic models, in general) is that not only id gives a forecast, but also the estimated probability associated with each possible outcome. Such information is crucial for management decision-making and is used for the Brier score performance measure (see below).

The most common measure of performance estimation in classification is accuracy, it measures the ratio of correct forecasts. However, accuracy ratio could be improved easily by using classifiers where the recruitment intervals are not balanced, i.e. one interval contains most of the data, then classifying all the cases within this interval (accuracy paradox). This leads to classifiers with high accuracy, but not useful models.

For this reason, it must be complemented with other performance measures that consider error distribution between all recruitment intervals, such as true positive rate (TP). TP is the rate of correctly classified cases for each recruitment interval, i.e. correct forecasts of recruitment in each interval to the total number of cases in each interval. Finally, the Brier score measure (Brier, 1950; van der Gaag and Renooij, 2001; Yeung et al., 2005) was calculated, as a complement to the 'accuracy' and the 'true positive rate per class'. Brier score for a set of events and their outcomes is the average deviation between the forecasted probabilities; thus, a lower score represents higher performance. Brier score can be considered as a calibration metric, which takes into account 'a posteriori' the probabilities assigned by the classifier to each possible outcome:

Table 5.4. Ranking of factors subsets selected with CFS multivariate factors subset selection method, using the leaving one out validation scheme (LOOCV) applied to anchovy and hake recruitment time-series, discretized at 3 intervals. Notes: (i) stability is the number of times that each subset of factors has been selected in the LOOCV process; (ii) the stability ('St.') of each variable set is being evaluated, instead of each variable stability itself; (iii) the end-user selected factors sets are shaded.

Multivariate factors selection using leaving one out validation for recruitment discretized in 3 intervals													
Anchovy recruitment index					Anchovy recruitment				Hake recruitment				
Expert		*Max_mean_tp*			*Max_accuracy*		*Max_mean_tp*		Expert		Max mean_tp & acc.		
Factors	St.	Factors	St.	Factors	St.	Factors	St.	Factors	St.	Factors	St.	Factors	St.
CLI1; UIBs_4502	16	CLI1; V_4503; UIBs_4502	28	CLI1; UIBs_4502	19	CLI1; Sunspot	18	TempAnom N; CLI2	18	TempAnom N; CLI2	8		
CLI1; UIBs_4502; TURB_4502	6	POL; CLI1; V_4503; UIBs_4502	9	POL; CLI1; UIBs_4502;	16	POL; CLI1; Sunspot	2	TempAnom A; TempAnom N; CLI2; TURB_4502	3	TempAnom N; CLI2; TURB_4502	7		
CLI1; NAOm; UIBs_4502	4	CLI1; V_4503; UIBs_4502	1	CLI1; V_4503; UILm_4502; UIBs_4502;	3	EA; CLI1; Sunspot	1	AMO (unsmoothed); TempAnom N; CLI2; TURB_4502	2	AMO (unsmoothed); TempAnom N; CLI2; TURB 4502	5		
CLI1; NAOm; UIBs_4502; TURB_4502	3	PNA; CLI1; V_4503; UIBs_4502	1	POL; CLI1; UILm_4502;	1	NAO; CLI1; Sunspot	1	AMO (unsmoothed); TempAnom N; UIBm_4502; TURB_4502; CLI2;	1	AMO (unsmoothed); TempAnom N; CLI2; TURB_4502; Sunspot	3		
TNH; CLI1; UIBs_4502	2							AMO (unsmoothed); TempAnom N; CLI2; CLI3; TURB_4502	1	AMO (unsmoothed); TempAnom N; CLI2; UIBm_4502; TURB_4502	1		
POL; CLI1; UIBs_4502	1							AMO (unsmoot.); TempAn N; CLI2; CLI3; TURB_4502	1	TempAnom N; CLI2; EW (53.5N; 12.5W)	1		
CLI1; Uim_4311;	1							CLI2 UILm_4502; UIBs_4502	1	AMO (unsmoothed); TempAnom N; CLI2; CLI3; TURB_4502	1		
CLI1; UILm_4502; UIBs_4502	1							SSB; TempAnom N; CLI2; TURB_4502; NS (46.5N; 4.5W)	1	CLI2	1		
POL; CLI1; NAOm; UIBs_4502	1							TempAnom A; TempAnom N; CLI2; UILm_4502; TURB_4502; NS (46.5N; 4.5W)	1	SSB; TempAnom N; CLI2; TURB_4502	1		
CLI1; NAOm; V_4503; UIBs_4502	1							TempAnom N; CLI2; TURB_4502; NS (46.5N; 4.5W)	1	Central England temp; TempAn E; TempAn N; CLI2; TURB_4502	1		
CLI1; NAOm UIBs_4502; AA_Ind.	1												

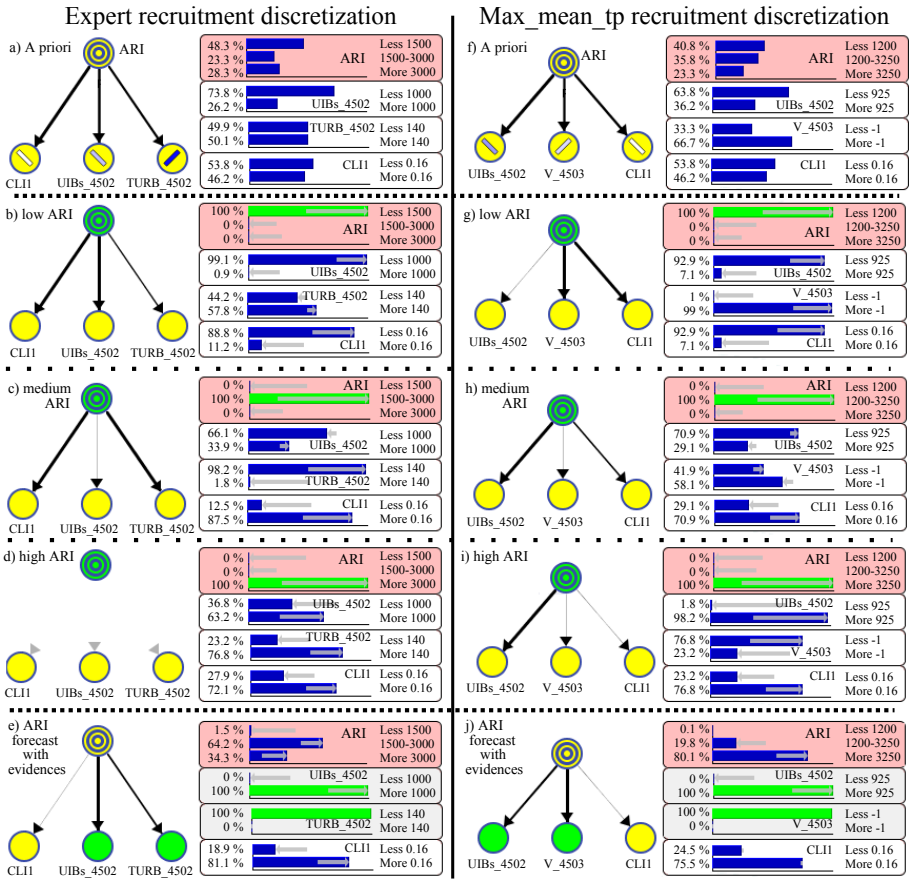


Fig. 5.5. Different ARI scenarios for 3 levels of recruitment, by the end-user discretization (first column) and by *max_mean_tp* ARI discretization (second column). Both figures enclosed in the first row show the 'a priori' probabilities of recruitment and factors. The figures enclosed in the second to fourth rows show the scenarios for low, medium and high recruitment. The figures enclosed in the last row shows the forecast for year 2008. In the left, the forecast provided by the evidence of upwelling (UIBs_4502) greater than 1000 and turbulence (TURB_4502) less than 140 for the expert recruitment discretization. In the right, the forecast provided by the evidence of upwelling greater than 925 or wind (V_4503) less than -1 for the *max_mean_tp* discretization strategy. The thickness of the arc is proportional to the strength of the probabilistic relationship it represents.

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{r}(p_l^k - y_l^k)^2$$

where N is the number of cases, r is the number of class values, and p_l^k is the forecasted probability for the l^{th} class value for the k^{th} case. The y_l^k value is 1 if l is the observed (correct) value of the class and 0 otherwise.

5.1.4 Methodology validation

A global and robust validation procedure is necessary, in order to establish the reliability of the methodology and final forecasts, as well as for comparison of classifiers (Bouckaert and Frank, 2004). Honest performance estimation requires the separation into training and test data, as a prior step to all the model-building, not just before classifier learning (Reunanen, 2003; Statnikov et al., 2005). Test instances must be retained totally unseen, whilst only the training data are used in all of the steps (supervised discretization, factor selection and model learning).

This approach permits the avoidance of over-fitting, that leads to over-optimistic and unstable performance estimations. In order to obtain robust estimations, a 10 times repeated 5-fold cross-validation (10x5cv; Fig. 5.3) scheme has been used, as recommended in Bouckaert and Frank (2004). In addition, the proposed wrapper recruitment discretization is guided by performance measurements. Thus, an inner loop with 100 re-sampling sets in a 0.632*bootstrap* schema (different from the outer loop) for calculating this performance is necessary (see Table 5.3). In order to ensure honest validation, the test data in the outer loop (Fig. 5.3) is not used in the inner loop (model-building in a bootstrapping scheme), where the discretization is being validated.

All of the above steps have been implemented using Weka API machine-learning software (Witten and Frank, 2005). Reproducibility is ensured by a Java programming language implementation of all the methodology, available from the ISG group webpage (*www.sc.ehu.es/ccwbayes/members/jafernandes* or at *www.azti.es*). In addition, Matlab (*www.mathworks.com*) and R (*www.r-project.org*) has been used for time-series analysis (Fig. 5.2). Finally, Bayesia software (*www.bayesia.com*) has been used for the visualization of the *naive Bayes classifier* (Fig. 5.5).

5.2 Results

Discretization:

Table 5.3 presents an example of the Anchovy Recruitment Index (ARI) discretization cut-off points proposed by the method together with their estimated performance and selected factors. Tables 5.5 and 5.6 present the different metrics used to evaluate the cut-off points, established using different validation methods for different recruitment discretization strategies.

Table 5.5. Performance evaluation of discretization methods, for different number of anchovy recruitment index intervals (bins). Notes: (i) several performance estimations are presented, to illustrate the differences in reported accuracies, depending upon the validation scheme; (ii) the scheme used in this work is 10 times 5-fold cross-validation (10x5cv); (iii) 'best 5cv' is the accuracy that would be reported, if the best repeat, is selected. 'Best fold' represents the best accuracy of a single fold in a specific cross-validation repeat (a 80-20% hold-out); (iv) true positive rates (TP) are followed by the recruitment cut-off point sets and the number of instances in each interval; (v) the classifiers that have been considered, as useful candidates to be used as the final classifier, are shaded.

Bins	Metrics	Equal frecuency	Equal width	Expert	*Max_mean_tp*	*Max_accuracy*
2	10x5cv Acc.	$73.7 \pm 4.9\%$	$81.9 \pm 5.2\%$	$71.2 \pm 3.9\%$	$65.1 \pm 5.5\%$	$67.4 \pm 4.9\%$
	Best 5cv Acc.	$82.1 \pm 18.9\%$	$90 \pm 16.3\%$	$76.8 \pm 18.7\%$	$74.3 \pm 9\%$	$76.8 \pm 16.5\%$
	Best fold Acc.	100%	100%	100%	87.5%	100%
	Brier score	0.08 ± 0.2	0.03 ± 0.02	0.08 ± 0.03	0.19 ± 0.05	0.10 ± 0.07
	TP low	79.9%(< 1550; 20)	93.5%(< 3600; 33)	77.1%(< 1500; 21)	61.8%(> 1050; 14)	74.4%(< 3600; 33)
	TP high	67.4%(> 1550; 19)	14%(> 3600; 6)	67.7%(> 1500; 17)	54%(> 1050; 25)	28.4%(> 3600; 6)
3	10x5cv Acc.	$41.3 \pm 9.2\%$	$62.4 \pm 3.6\%$	$47.4 \pm 7.1\%$	$44.9 \pm 5\%$	$47.1 \pm 7.6\%$
	Best 5cv Acc.	$53.9 \pm 10.5\%$	$69.6 \pm 20.6\%$	$55.7 \pm 21.5\%$	$51.4 \pm 20\%$	$58.9 \pm 10.4\%$
	Best fold Acc.	75%	87.5%	100%	75%	75%
	Brier score	0.21 ± 0.05	0.10 ± 0.04	0.16 ± 0.03	0.24 ± 0.05	0.23 ± 0.04
	TP low	47.1%(< 1000; 13)	86.4%(< 2500; 27)	75.6%(< 1500; 19)	47.3%(< 1200; 16)	50.4%(< 1500; 19)
	TP medium	32.9%(10002400; 13)	9.2%(25005000; 10)	24.3%(15003000; 9)	27%(12003250; 14)	24.4%(15003250; 11)
	TP high	51.8%(> 2400; 13)	2%(> 5000; 2)	28.1%(> 3000; 11)	39.4%(> 3250; 9)	41%(> 3250; 9)
4	10x5cv Acc.	$33.4 \pm 6.3\%$	$57.1 \pm 4.3\%$	–	$30.8 \pm 4.1\%$	$26.93 \pm 6.8\%$
	Best 5cv Acc.	$41.4 \pm 12.5\%$	$64.3 \pm 9.7\%$		$36.4 \pm 18.1\%$	$38.2 \pm 11.7\%$
	Best fold Acc.	62.5%	87.5%		62.5%	50%
	Brier score	0.25 ± 0.04	0.13 ± 0.02		0.34 ± 0.06	0.31 ± 0.04
	TP low	49.7%(< 850; 10)	84.1%(< 2000; 23)		36.5%(< 1050; 14)	43.3%(< 1050; 14)
	TP med. I	10%(8501550; 10)	18.7%(20003700; 10)		10.8%(10501900; 9)	11.3%(10501900; 9)
	TP med. II	27.7%(15503250; 10)	17%(37005400; 5)		15%(19003350; 9)	11.3%(19003350; 9)
	TP high	51.8%(> 3250; 9)	0%(> 5400; 1)		35.7%(3350 >; 8)	30.4%(> 3350; 8)

The fisheries expert considered that three intervals (low, medium and high recruitment) were the most useful for management purposes and proposed 1500 and 3000 as cut-off points for ARI (Table 5.5) together with 170000 and 220000 for hake recruitment(HR; Table 5.6). The *max_mean_tp* recruitment discretization method has suggested similar cut-off point sets: 1200 and 3250 for ARI (Table 5.5); and for HR, 170000 and 213000 (Table 5.5). The expert ARI discretization does not accomplish significantly higher accuracy (47.4%) compared to the *max_mean_tp* discretization (44.9%; p>0.05, *corrected paired t-test*; Table 5.5). Expert boundaries show a lower Brier score (0.16) performance than *max_mean_tp* (0.24). However, the true positive distribution between the different intervals defined by the expert is imbalanced. Therefore, the classifier using expert boundaries shows a good forecasting behaviour in low recruitments (75.6%), but a poor behaviour for the rest of recruitments levels (medium 24.3% and high 28.1%). The *max_mean_tp* discretization method shows a more balanced distribution of the error (47.3% low, 27% medium and 39.4% high).

In the case of hake, fisheries expert and *max_mean_tp* discretizations accomplish similar accuracies (42.8% compared to 42.2%; Table 5.5), with no significant differences (p>0.05; corrected paired t-test). Expert levels boundary set presents lower Brier score (0.23) than *max_mean_tp* (0.30). The discretization provided by the experts shows again an imbalanced distribution of performance between intervals, showing a good forecasting behaviour in low recruitments (71.7%) and poor behaviour in high recruitment, whereas *max_mean_tp* shows a more balanced distribution of true positive rates (41.3% low, 33.6% medium and 45% high). Taking into account both cases, the *max_mean_tp* semi-automated discretization performed satisfactorily, helping to avoid overfitting, as well as establishing informative recruitment intervals (similar to those proposed by the expert) and with a balanced distribution of the error among all the intervals.

Factors selection:

After discarding unbalanced recruitment discretizations and classifiers with unbalanced true positive distributions, all shaded classifiers in Tables 5.5 and 5.6 are selected as potential candidates to be used as the final classifiers. Two interval recruitment discretizations accomplish 70% of accuracy; and Brier scores of around 0.10 for Anchovy Recruitment Index and Hake Recruitment. Whereas, three interval recruitment discretizations accomplish 50% of accuracy; and Brier scores of around 0.25 for ARI, and 0.30 for HR.

Regarding the multivariate factors selection results for ARI (Table 5.4), the most stable subset of factors for end-user discretization was formed by CLI1 (the first PCA component of climatic detrended indices, being the most influential EA/WR and POL indices) and UIBs_4502 (Upwelling Index along the French and Spanish coasts (45N, 2W) annual mean of positive values, March-July, $m^3 s^{-1} km^{-1}$). The same subset of factors is the second top-ranked

Table 5.6. Performance evaluation of discretization methods, for different number of hake recruitment index intervals (bins). Notes: (i) several performance estimations are presented, to illustrate the differences in reported accuracies, depending upon the validation scheme; (ii) the scheme used in this work is 10 times 5-fold cross-validation (105cv); (iii) 'best 5cv' is the accuracy that would be reported, if the best repeat is selected. 'Best fold' represents the best accuracy of a single fold in a specific cross-validation repeat (a 80-20% hold-out); (iv) true positive rates (TP) are followed by the recruitment cut-off point sets and the number of instances in each interval; (v) the classifiers that have been considered, as useful candidates to be used as the final classifier, are shaded.

Bins	Metrics	Expert	*Max_mean_tp*	*Max_accuracy*
2	10x5cv Acc.		$68.3 \pm 8.2\%$	$68.5 \pm 6.6\%$
	Best 5cv Acc.		$79.3 \pm 18.2\%$	$76 \pm 8.6\%$
	Best fold Acc.	–	100%	100%
	Brier score		0.12 ± 0.08	0.10 ± 0.07
	TP low		$54.1\%(< 170k; 7)$	$63.9\%(170k; 7)$
	TP high		$67\%(> 170k; 22)$	$55.6\%(170k; 22)$
3	10x5cv Acc.	$55.7 \pm 6.4\%$	$43.7 \pm 7.5\%$	$43.3 \pm 7\%$
	Best 5cv Acc.	$66 \pm 10.7\%$	$52 \pm 17.3\%$	$54.6 \pm 20.4\%$
	Best fold Acc.	100%	83.3%	83.3%
	Brier score	0.23 ± 0.05	0.30 ± 0.05	0.32 ± 0.05
	TP low	$71.7\%(< 170k; 7)$	$41.3\%(< 170k; 7)$	$43.7\%(< 170k; 7)$
	TP medium	$57.9\%(170220k; 16)$	$33.6\%(170213k; 14)$	$36.3\%(170213k; 14)$
	TP high	$17\%(> 220k; 6)$	$45\%(> 213k; 8)$	$38\%(> 213k; 8)$
4	10x5cv Acc.		$33.6 \pm 8.3\%$	$32 \pm 7.7\%$
	Best 5cv Acc.		$44.7 \pm 18.4\%$	$41.3 \pm 14.5\%$
	Best fold Acc.		66.7%	66.7%
	Brier score	–	0.38 ± 0.04	0.40 ± 0.05
	TP low		$42\%(< 161k; 6)$	$47.3\%(< 161k; 6)$
	TP med. I		$11\%(161205k; 9)$	$5\%(161197k; 5)$
	TP med. II		$26.7\%(205247k; 9)$	$25.7\%(197213k; 10)$
	TP high		$35\%(> 247k; 5)$	$39.7\%(> 213k; 8)$

set of variables for the end-user discretization, with an additional factor; TURB_4502 (mean annual turbulence Bay of Biscay at 45N, 2W, $m^3 s^{-3}$). For *max_mean_tp* discretization method, UIBs_4502, V_4503 (mean N-S wind; 45N, 03W, ms^{-1}) and CLI1 were the selected factors. However, *max_mean_tp* recruitment discretization shows stronger probabilistic relationships in the final classifier graphical representation (Fig. 5.5), coupled with a greater factor subset stability (Table 5.4). Other factors that emerge in the multivariate factors selection ranking, with a lower level of stability than the exposed ones

(Table 5.4), are the following: NAOm (North Atlantic Oscillation annual mean from Hurrell, 1995); the climatic indices TNH and POL; and UILm_4502 (Upwelling Index Landes 45N, 2W annual mean, $m^3 s^{-1} km^{-1}$).

The end-user, the *max_mean_tp* and *max_accuracy* discretization methods select the same set of factors, in the HR case (Table 5.4). This set is composed of: TempAnom N (mean annual temperature anomaly for the area 55-60N; 15-10W); CLI2 (second PCA component of de-trended climatic indices, with the most influential being the NAO index); TURB_4502; and Sunspot (number of Sunspots per year).

Performance and recruitment scenarios:

Tables 5.5 and 5.6 present the performance of the different recruitment discretization methods measured in terms of several performance indicators and validation schemes. Table 5.7 lists a set of examples of the estimated probability associated with each forecast, which should contribute to management decisions.

Table 5.7. Forecast output from a cross-validation fold, for 10 test samples or years. Notes: (i) the first column is the labelled recruitment and the second column is the label forecasted by the *naive Bayes* classifier; (ii) an incorrectly classified instance is marked with the symbol '+' in the third column; (iii) the forth to sixth columns are the 'a posteriori' probabilities for each recruitment level, after observing the factors; (iv) an example of a misclassified case is shadowed. '*' indicates the most probable recruitment forecast value, for a given year.

Observed	Forecasted	Error	Less_1500	1500-3000	More_3000
More_3000	More_3000		0.005	0.065	0.931_*
1500-3000	More_3000	+	0.119	0.376	0.505_*
More_3000	More_3000		0.070	0.132	0.798_*
Less_1500	Less_1500		0.715_*	0.029	0.257
More_3000	More_3000		0.001	0.010	0.989_*
1500-3000	1500-3000		0.025	0.958_*	0.016
Less_1500	Less_1500		0.517_*	0.207	0.276
Less_1500	Less_1500		0.931_*	0.046	0.023
Less_1500	Less_1500		0.960_*	0.035	0.005
1500-3000	1500-3000		0.220	0.581_*	0.200

Finally, different scenarios for anchovy can be observed for the *naive Bayes classifier* (Fig. 5.5). The scenarios for the ARI *max_mean_tp* discretization method are the following: (i) in the 'a priori' recruitment distribution (Fig. 5.5f), Low Recruitment is the most predominant level (41%), followed by Medium (36%) and High Recruitment (23%); (ii) a Low Recruitment level (Fig. 5.5g) is characterised by high N-S wind (V_4503) and low CLI1; (iii)

Medium Recruitment (Fig. 5.5h) is characterised by a probable high CLI1, dominated slightly by low N-S wind and a probable high upwelling; and (iv) High recruitment (Fig. 5.5i) is characterised by high upwelling episodes.

Neither in the case of Anchovy Recruitment (AR time-series shorter than ARI, Table 5.5) or Hake Recruitment (HR, Table 5.6) was the spawning biomass selected by the model, as a relevant factor. This outcome is not surprising if it is considered that, in both cases, the spawning stock biomass explains very little of the recruitment variability (Fig. 5.1.1). However, it has to be considered that in the case of anchovy, the recruitment time-series (AR) is probably too short (21 years) to adopt the approach described. In hake time-series there is an absence of data on very low recruitment values.

5.3 Discussion

Stock-recruitment models (Ricker, 1954; Beverton and Holt, 1957), together with other regression methods used for environmental variables (Schirripa and Colbert, 2006; Planque and Buffaz, 2008), are used commonly for recruitment forecast. A complementary methodology is presented here, that converts the regression problem of recruitment forecast (quantitative), into a classification problem (qualitative), that is less complex (Alpaydin, 2004; Bishop, 2006). Thus, the variable of interest (recruitment) needs to be discretized.

A notable advantage of the proposed approach is that certain classification models, such as *naive Bayes*, provide the probabilities associated to each output or 'recruitment interval'; these can be used as a measure of robustness of the forecast (Frank et al., 2000). In terms of management, instead of trying to forecast a number with high precision (i.e. regression provides an exact number within a problem with high *uncertainty*), an interval with a measure of the *uncertainty* of that forecast can be a more useful outcome. Different interpretations and decisions will be made if a forecast shows a probability of 0.50 or 0.9 (Table 5.7). Other differences with respect to regression-based models are the following: no strong distribution assumption is made; and it allows to deal with a certain degree of missing data to be managed, with the outliers and scale effect reduced (Witten and Frank, 2005).

Regression-based models attempt to fit a model to the data by likelihood maximisation; whilst some classification models have an additional aim to the likelihood that is the forecast performance (Zhou, 2003). Direct comparison between regression and classification approaches is not possible quantitatively as they use different fitting and performance measures. However, because stock-recruitment relationships often fail (Myers et al., 1995) and relations with the environment are difficult to disentangle, managers have to work with recruitment scenarios (Schirripa and Colbert, 2006; Planque and Buffaz, 2008). In such a situation, the interest of having a forecast and a measure of the *uncertainty* associated with each forecast is important.

The major contribution of this study is the proposal of a pipeline of already established methods, in order to ensure a robust recruitment forecast, based upon scarce and noisy data. Once robustness has been assured, a 'trade-off', between searching for a high accuracy degree with informed factors selection and over-fitting avoidance, can be used for domain knowledge extraction (Fayyad et al., 1996).

5.3.1 Proposal for a robust supervised classification pipeline

Discretization dependency:

The supervised discretization and selection of variables have advantages in model-building and for expert interpretation (Fayyad et al., 1996). This means that supervised steps are dependant upon the recruitment definition (interval boundaries). This is the reason why different recruitment discretizations can lead to different factor subsets and the motivation for suggesting a semi-automatic method to identify adequate recruitment boundaries (Uusitalo, 2007). Among the proposed boundaries that improve performance, the end-users can employ their expertise to choose cut-off points that have biological meaning. As an example, in Figure 5.5, the factors selected with the cut-off points proposed by the fisheries expert do not show a good behaviour for forecasting of high recruitment level. In contrast, those cut-off points selected by the proposed *max_mean_tp* discretization method, show a competitive behaviour in all the recruitment intervals and they are close to the expert proposed cut-off points. Equal width discretization leads to extremely unbalanced recruitment levels, which are not particularly useful. Equal frequency discretization leads to artificial boundaries, without any biological meaning; thus, these are not useful for subsequent interpretation by the end-users. The proposed semi-automated recruitment discretization method considers critical issues in order to learn a *Bayesian network* (Uusitalo, 2007): the number of intervals or bins (Table 5.5 and 5.6); the domain significance of the breakpoints (through entropy reduction and end-user evaluation; Table 5.1); and the balance of the instances in each recruitment level using a threshold (a minimum number of instances within each interval).

As the number of re-sampling sets in the bootstrapping scheme is increased, both (*max_mean_tp* and *max_accuracy*) wrapper recruitment discretization criteria suggest similar cut-off point sets. At least 100 bootstrap re-sampling sets are suggested, to ensure robustness; if lower number of re-sampling sets is performed, our suggestion is to use the *max_mean_tp* criterion. Significance is 're-ensured' in factor's discretization by means of a well-known entropy-based supervised discretization algorithm, such as the Fayyad & Irani Multi Interval Discretization method (Fayyad and Irani, 1993). Even if accuracy does not increase significantly, forecasts benefit from a more informed model (Guyon et al., 2007). Finally, fisheries experts found useful a method that allows them to find recruitment and factors intervals that are reciprocally

connected and that can have biological meaning. This is extremely useful for interpretation and knowledge extraction.

Multivariate non-redundant variable selection:

The Correlation-based Feature Selection method (CFS) tends to select non-redundant factors that are likely to be independent of each other (Hall and Smith, 1997). This approach favours classifiers such as *naive Bayes*, which restricts relationships between factors and takes advantage of variables which are independent of each other, given the recruitment. CFS often selects the top ranked factors, in a univariate ranking of correlations with the recruitment, together with other factors that are not the top ranked, but that are selected due to their non-redundant nature. The selection of not top ranked factors is related to their forecast power in intervals, not discriminated by other variables that are ranked higher and redundant with other top ranked variables. The properties of this CFS scheme makes the returned subset of factors interesting, for expert discussion.

Naive Bayes classifier:

There are several specific properties of *naive Bayes* classifier (Domingos and Pazzani, 1997; Zhang, 2004) that makes them especially useful for the proposed methodology, together with its objectives. Firstly, its probabilistic nature is particularly useful for management decisions, where information on the estimated probability of each outcome can be decisive (Table 5.7). The end-user can check if the model forecast is strong enough (high 'a posteriori' probability for the most likely recruitment level), or if there are several forecast with similar 'a posteriori' probabilities. These equally probable situations occur usually when the recruitment is close to the boundaries of two recruitment levels (Frank et al., 2000). Secondly, it benefits from techniques such as CFS and PCA (Principal Components Analysis) that return non-redundant factors.

Finally, no other classification method has a performance that is significantly higher (*corrected paired t-test*) than *naive Bayes* in this study (Table 5.8). Indeed, *naive Bayes* shows higher performance estimation stability (lower variability between repeated cross-validations). *Support vector machines* accomplish similar results to *naive Bayes*. It is important to highlight that for other recruitment time-series, another classification method can outperforms the *naive Bayes* classifier. If a longer recruitment time-series is considered a *Tree Augmented Naive Bayes* (TAN) might benefit from interactions between factors and outperform a *naive Bayes classifier*. Finally, it might not be justified the use of a more complex or less comprehensible model by a small performance improvement following Occam's razor principle (Domingos, 1999).

Table 5.8. Comparison of *naive Bayes* with other classification models as the classifier in the proposed pipeline of supervised methods. The CPU-time rows correspond with the processing time required to perform the recruitment *max_mean_tp* discretization. The methodology validation consists of 10 times-repeated 5-fold cross-validation with a inner 100 re-sampling sets in a *0.632 bootstrap* schema. Classification models abbreviations are the following: NB for *naive Bayes*; TAN for *tree augmented naive Bayes*; J48DT for *J48 Decision Tree*; MPNN for *multi-layer perceptron neural network*; SVM for *support vector machine*.

Species	Metrics	NB	TAN	J48DT	MPNN	SVM
ARI	10 x 5cv Acc. (%)	44.9 ± 5	38.4 ± 9.1	46.3 ± 7.3	46.3 ± 7.7	45.8 ± 5.1
	Brier score	0.24 ± 0.05	0.26 ± 0.06	0.27 ± 0.05	0.29 ± 0.05	0.22 ± 0.05
	TP low	0.473	0.393	0.488	0.474	0.454
	TP medium	0.27	0.276	0.313	0.29	0.376
	TP high	0.394	0.323	0.348	0.356	0.325
	CPU-time (min)	29	29.8	29.7	82.3	33.4
HR	10 x 5cv Acc. (%)	43.7 ± 7.5	32.9 ± 7.6	41.5 ± 7.7	44.6 ± 7.5	44.6 ± 9.4
	Brier score	0.30 ± 0.05	0.38 ± 0.06	0.37 ± 0.07	0.36 ± 0.07	0.26 ± 0.04
	TP low	0.413	0.32	0.357	0.42	0.43
	TP medium	0.336	0.342	0.366	0.379	0.316
	TP high	0.45	0.189	0.358	0.364	0.403
	CPU-time (min)	10.6	10.7	10.6	24.1	12.4

Methodology validation:

Repeated cross-validation reduces the variance of the estimated performance of the classifier (Rodríguez et al., 2010). In addition, repeated cross-validation permits an analysis of the stability of the classifier performance, reporting replicable as well as cautious performances (Bouckaert and Frank, 2004). As such, it can be observed that single hold-out validation could reach 100% accuracy (Tables 5.4 and 5.6). This conclusion is not realistic and can undermine the trust of the managers that have to rely on these performance estimations.

Cross-validation provides safer estimations, reporting a more cautious performance (mean of internal cross-validation accuracies) and stability (standard deviation). In this study datasets, the use of repeated cross-validation reported lower accuracies, up to 10% less than the cross-validation (Tables 5.4 and 5.6). However, it reports also lower deviations. A high estimated accuracy of 70%, with a high variance of 20%, is not useful. It is preferable to obtain a lower accuracy value with a small variance. Finally, in order to ensure replicability of the results, repeated cross-validation can be used for methodology comparison, through a statistical test.

The inclusion of all the steps in the validation procedure (i.e. train-test split before discretization, factor selection and model learning), avoids over-

fitting (Reunanen, 2003; Statnikov et al., 2005). In addition, it ensures robustness of the reported performances, selected factors and discretization intervals. Therefore, the reported results would be less sensitive to new data, or to any changes in the available data.

5.3.2 Selected factors

A way to asses the validity of the method is to check whether the selected factors make sense, compared with the published literature and under the view of several fisheries experts.

For the ARI, the factors selected in the multivariate and non-redundant CFS method (for both, end-user and semi-automated recruitment discretizations) are the CLI1 global pattern index, upwelling index, turbulence and wind speed (CLI1, UIBs_4502, TURB_4502 and V_4503). The first component of the PCA on climate indices (CLI1) reflects variations of the Eastern Atlantic pattern. The selected factors coincide with previous knowledge showing the effect of the Eastern Atlantic pattern on Anchovy Recruitment (Borja et al., 2008), upwelling intensity (Borja et al., 1998; Allain et al., 2001), turbulence (Allain et al., 2001) and wind-driven offshore transport (Irigoien et al., 2007).

In general, the ARI time-series analysis reveals a periodical component, which is similar to the CLI1 behaviour (Fig. 5.2b); this is in accordance with Bode et al. (2006), who found also a coincidence between anchovy landings and CLI1. Such cycles in both variables explain the high forecast power of the CLI1. However, the time-series in the dataset used in the present study is not sufficiently long to confirm such cyclic behaviour. Moreover, it can be observed in Fig. 5.5, that CLI1 is especially relevant, for low and medium recruitments.

In the case of hake, the factors of higher correlation with recruitment (that have been selected in the multivariate and non-redundant CFS) are CLI2, temperature anomaly and turbulence. In addition, Sunspot number has been selected, even if it is not amongst the highest ranked variables in the univariate ranking of factor-recruitment correlation values. The second component of the PCA on climate indices (CLI2) reflects mainly the variability in the Northern Atlantic Oscillation (NAO), the first prominent mode of low-frequency variability over the North Atlantic that modulates the temperature and precipitation regime (Hurrell et al., 2003). Meiners (2007) has demonstrated already a robust and persistent influence of the NAO, upon the recruit abundance of hake, off the north western coast of Africa. In this way, a positive relationship between NAO and the extent of the environmentally-optimal window for recruitment was found, through longer periods of wind-induced upwelling episodes.

Turbulence episodes over the hake spawning area, i.e. shelf break (Álvarez et al., 2004), represented by TURB_4502, would have a major role in egg and larval survival, through their food encounter and capture probability as shown for other gadoids (e.g Fiksen et al. (1998)).

Temperature is related not only to NAO, but appears also to be related to solar activity (Benner, 1999). Within this context, sea-surface temperature (SST) fluctuations are believed to be driven by, partly, Sunspot cycles. Reid (1987, 1991) found a remarkable similarity between SST anomalies and the 11-year running mean of the Sunspot number. Since solar variability is believed to play a prominent role in recent global temperature change (Lean et al., 1995), whilst HRI has been found to be related to SST anomalies, Sunspot number influence on the Hake Recruitment can be understood (through the effect of SST anomalies on hake).

Amongst different temperature anomaly measurement boxes, Temperature anomaly N is the only one selected as an HRI factor in the multivariate and non-redundant CFS final selection. It represents the average temperature anomaly over the study period, within the area located around 55-60N, 10-15W. This area is known as the Rockall Trough and lies close to some of the main recruitment areas of hake, over the shelf break of the Celtic Sea (Ibaibarriaga et al., 2007). The Rockall Trough provides a pathway by which warm North Atlantic upper water reaches the Norwegian Sea. In this sense, the effect of the temperature anomaly, on the Rockall Trough area can be understood as an increased transport of warm water, with eggs and larvae, to northern areas. There could be a consequent increase in the recruitment areas and a direct beneficial effect of the warmer temperatures, for the eggs and larvae spawn at the northern limits of the hake distribution.

It can be observed that neither in the case of Anchovy Recruitment nor in the case of Hake Recruitment (AR and HR) is the spawning stock biomass (SSB) selected as a factor. This outcome is easily understandable, when it can be observed that, in both cases, SSB explains only a small amount of the recruitment variance: low, medium and high recruitments can be found at the same SSB levels (Fig. 5.1.1). However, the case of anchovy has to be considered with caution, as the AR time-series is probably too short for accurate results and SSB was not available for the longer recruitment index time-series (ARI). The use of SSB and other environmental variables, as factors, together with their interactions have been discussed widely in the literature (e.g. Schirripa and Colbert (2006); Planque and Buffaz (2008)).

The proposed methodology properties are interesting for data analysis. However, it has some limitations that, like any other statistical tool, must be taken into consideration by the users. The key factors selected depend on the user assumptions, on what is or could be relevant for recruitment, as well as the selected recruitment boundaries (even using the *max_mean_tp* method, the user selects the boundaries from a ranked list).

Therefore, pre-existing knowledge and common sense must guide the analysis. Despite the analysis being effective in removing spurious correlations and identifying sets of complementary factors, there are some scenarios where a spurious correlation can survive this analysis, as for example: i) if the end-user has restricted the candidate set of variables to highly correlated variables and has not provided any good complementary variables; ii) variables are highly

correlated by chance, e.g. recruitment time-series with a negative trend (overfishing) and temperature always shows a positive tendency (climate warming); iii) variables that could explain just a single recruitment interval, not explained by other variables.

5.4 Conclusions and suggestions for Future Work

The proposed methodology (Fernandes et al., 2010c) permits a robust classifier learning procedure, ensuring stable results, i.e. results that do not dramatically change with slight changes in the data. This outcome is accomplished by the use of well-established methods in the machine-learning literature which properties are known, as well as using strong and honest validation in all of the steps and over the whole model-building process.

Firstly, the recruitment discretization algorithm helps the expert to identify more informed and stable boundaries, which is validated using a bootstrapping re-sampling procedure for over-fitting avoidance. Secondly, the Multi Interval Discretization algorithm identifies factor boundaries that are significant for recruitment forecast, avoiding over-fitting by using a minimum description length criteria. Thirdly, the CFS method for factors selection identifies a non-redundant and more stable set of factors by means of a 'leaving one out' re-sampling process. Fourthly, the model performance estimation is undertaken by validating not only the model, but also the rest of the data analysis steps for a honest validation. Validation that consists in repeated cross-validation in order to report reliable and reproducible estimated performances.

All of the above conclusions, together with the fact that the model estimates the *uncertainty* inherent in the reported forecast, make the final model robust and reliable even with sparse data. Hence, the proposed methodology is a valid alternative when traditional methods, based upon stock-recruitment relationships, cannot be applied. It may be of particular interest in the context of forecasting as it fits with the concept of projection according to weighted mixture of past recruitment (years or levels). It can be also useful as an exploratory analysis when the objective is to understand the factors determining recruitment, or as a second opinion to other models forecasts.

There are several issues to explore for future work: (i) consideration of temporal relationships, between factors and the recruitment; (ii) consideration of relationships with other species recruitment, due to spatial and food competition or predation; and (iii) the use of cost-sensitive classification and model-building, where all kinds of errors are not equally penalised.

5.5 Posterior work and robustness through time

Robustness of the methodology can be observed comparing selected factors for anchovy recruitment in a later study (Fernandes et al., 2009b) and selected

factors in the first study (Fernandes et al., 2010c), which corresponds to the previous exposed results. Between both studies several modifications to the database have been performed: the anchovy recruitment time-series has been recalculated, more factor candidates has been added and some of the factors have been removed because they are not published anymore.

However, the same factors has been selected (upwelling and CLI1) or they have been replaced by a similar one when eliminated from the analysis. This is the case of the eliminated 'ICOADS N-S wind annual mean (45N, 03W)' that has been replaced by the 'FNMOC N-S wind stress annual mean (45N, 02W)', which was not considered in the previous work. Similarly, in the case of hake, TempAnom N (mean annual temperature anomaly for the area: 55-60N, 15-10W) is selected again. However, there are some discrepancies, mainly due to the inclusion of factor candidates that where not considered in the previous work like Ekman transport or wind data. Number of Sunspots is replaced by sun geomagnetic activity, both related between them and with sea temperature (Reid, 1987). Other temperature related factors are selected such as Global Tanom and SST; whose might be influencing copepod abundance (Hays et al., 2005).

The expose research has been published and leaded to the following research contributions in refereed journals and international forums:

- (2010) **Fish recruitment prediction, using robust supervised classification methods**. *Fernandes J.A.*, Irigoien X., Goikoetxea N., Lozano J.A., Inza I., Pérez A. and Bode A. Ecological Modelling, 221(2): 338-352.
- (2010) **Robust machine-learning techniques for recruitment forecasting of North East Atlantic fish species**. *Fernandes J.A.*, Irigoien X., Lozano J.A., Inza I. and Pérez A. ICES Journal of Marine Science. Submited.
- (2010) **The potential use of a Gadget model to forecast stock responses to climate change in combination with Bayesian Networks: the case of the Bay of Biscay anchovy**. Andonegi E., *Fernandes J.A.*, Quincoces I., Uriarte A., Pérez A., Howell D. and Stefansson G. ICES Journal of Marine Science. Submited.
- (2009) **Robust approaches to supervised machine learning techniques for seven fish species recruitment prediction in fisheries**. *Fernandes J.A.*, Irigoien X., Goikoetxea N., Uriarte A., Lozano J.A. and Inza I. ICES/PICES/UNCOVER Symposium 2009 on Rebuilding Depleted Fish Stocks - Biology, Ecology, Social Science and Management Strategies. Warnemnde/Rostock (Germany). November 3-6.
- (2010) **The potential use of a Gadget model to forecast stock responses to climate change in combination with Bayesian Networks: the case of the Bay of Biscay anchovy**. Andonegi E., *Fernandes J.A.*, Quincoces I., Irigoien X., Pérez A., Howell D. and Stefánsson G. International Symposium Climate Change Effects on Fish and Fish-

eries: Forecasting Impacts, Assessing Ecosystem Responses, and Evaluating Management Strategies. Sendai (Japan). April 26-29th.

- (2010) **Use of juvenile abundance indices for the management of the Bay of Biscay**. Ibaibarriaga L., Uriarte A., Sanchez S., *Fernandes J. A.* and Irigoien X. Working document to WGANSA, 24-28 June 2010, Lisbon (Portugal).
- (2010) **UNCOVER: Fish stock recovery strategies - Report from the Bay of Biscay**. Andonegi E., Quincoces I., Murua H., *Fernandes J.A.*, Uriarte A., Sanchez S., Cerviño S., Velasco F., Huret M., Lehuta S. and Petitgas P.
- (2009) **Anchovy Recruitment Mixed Long Series prediction using supervised classification**. *Fernandes J.A.*, Irigoien X., Uriarte A., Ibaibarriaga L., Lozano J.A. and Inza I. Working document to the ICES benchmark workshop on short lived species (WKSHORT) Bergen (Norway), August 31st - September 4th.
- (2009-2010) Influence of the northeastern Atlantic oceano-meteorological variability on the northern hake (Merluccius merluccius), based on the last three-decadal period data (1978-2006). Goikoetxea N. PhD thesis. AZTI-Tecnalia.
- FACTs european project.
- UNCOVER european project.
- Theme 6 of the EC Seventh Framework program, through the Marine Ecosystem Evolution in a Changing Environment (MEECE No 212085) Collaborative Project.
- K-EGOKITZEN of the Basque Country Government.
- ECOANCHOA supported by the Department of Agriculture, Fisheries and Food of the Basque Country Government.

Finally, his work has appeared in Basque Country television through a collaborator in the science divulgative program *Teknopolis*.

6

Multi-dimensional fish recruitment forecasting

6.1 Introduction

In the previous chapter, classification methods have been presented as mono-species forecasting approaches (Fernandes et al., 2010c). Other studies can be found in the literature (Dreyfus-León and Chen, 2007; Dreyfus-León and Schweigert, 2008) where supervised classification has been applied. However, based in the ecosystem-based approach to fisheries management, it would be desirable to approach the multiple species fish recruitment forecasting simultaneously in a single classification model.

Classification models based upon probability theory, such as *Bayesian network* classifiers, are especially useful for fisheries management (Fernandes et al., 2010c). In addition, supervised pre-processing methods can be combined with *Bayesian networks* in order to improve classifier performance and interpretability (Fernandes et al., 2010c; Uusitalo, 2007). Data pre-processing is a key issue in a domain of high *uncertainty*, such as recruitment forecasting, where sparse and noisy data are common. Supervised pre-processing methods can also aid in the process of model interpretation and knowledge extraction (Fernandes et al., 2010c; Fayyad et al., 1996).

Classical classification methods can be applied to multiple species modelling using a different model for each class variable, or alternatively, using a single model where the class is a compound of all class variables by performing the Cartesian product. However, in small datasets, such as in the fish recruitment domain, this is not adequate because there will be many class values without data, or with too little data to be representative.

In addition, the use of this compound class reduces model readability and comprehensibility (van der Gaag and de Waal, 2006; de Waal and van der Gaag, 2007; Rodríguez and Lozano, 2008, 2010; Bielza et al., 2010). Therefore, the Cartesian product of class variables is not commonly used for multiple class-variables modelling. Consequently, in this work the utilization of separate models for each class variable is selected as uni-dimensional approach to

compare with the proposed multi-dimensional approach. In contrast, multi-dimensional *Bayesian networks* (MDBNs) permit the learning of classifiers that have multiple class variables in a single model. This approach is used in this study for multiple species modelling (Fernandes et al., 2010b). MDBNs have been proposed in (van der Gaag and de Waal, 2006); their learning and inference extended in (de Waal and van der Gaag, 2007) and a multi-objective learning approach has been developed in (Rodríguez and Lozano, 2008, 2010). In these studies, the multiple class variables approach is referred to as 'multi-dimensional'. This term is used also in this study, in order to avoid the longer term 'multiple class variables'.

As with one class variable classification (uni-dimensional), it would be desirable to be able to combine these MDBNs in a pipeline with specific pre-processing methods targeting several species forecasting, taking advantage of the relationships between species. Therefore, pre-processing supervised methods adaptation (missing data imputation, discretization and feature subset selection) is needed for the multi-dimensional approach.

Within this context, the objectives of this study are: i) to develop pre-processing strategies for multi-dimensional (Mul-D) classifiers based upon uni-dimensional (Uni-D) state-of-the-art methods; ii) to test the proposed pre-processing methods with synthetic datasets; and iii) to apply the proposed multi-dimensional approaches within the real domain of fish recruitment forecasting.

This choice of the *naive Bayes* classification model is motivated by two facts. On the one hand, *naive Bayes* for one class variable problems (uni-dimensional) has outperformed other more complex paradigms within the fish recruitment forecasting domain (Fernandes et al., 2010c), where data is usually scarce. On the other, as one of the objectives of this study is the design and evaluation of multi-dimensional pre-processing methods, the use of a fixed classifier permits to reduce the influence of the model learning process (Hua et al., 2009).

6.2 Performance measures for multi-dimensional classification

In order to evaluate the MDBN classifiers learnt, two commonly-used (uni-dimensional) performance measures have been generalized for the multi-dimensional approach: *accuracy* and *Brier score*.

In the uni-dimensional approach *accuracy*, or percent of correctly classified cases, measures model performance without considering the estimated 'a posteriori' probability of each class value. It considers only the class value with the highest probability (0-1 loss measure). *Accuracy* is measured between 0% and 100%, with the objective of the highest values that indicate the best results. *Brier score*, contrary to *accuracy*, considers the estimated 'a posteriori' probabilities for each possible outcome (Brier, 1950; van der Gaag et al., 2002;

Yeung et al., 2005). The lower the value of *Brier score* (between 0 and 2), the better the classifier:

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{r}(p_l^k - y_l^k)^2$$

where N is the number of cases, r is the number of class values, and p_l^k is the predicted probability for the l^{th} value of the class for the k^{th} case. The y_l^k value is 1 if l is the observed (correct) value of the class and 0 otherwise. In domains such as recruitment forecasting for fisheries management, the additional information provided by using the *Brier score* is valuable information for the decision-making process (Fernandes et al., 2010c).

The *accuracy* measure adapted to the multi-dimensional approach has been considered in two variants:

1. Using the average of the *accuracy* measure calculated for each class variable in isolation, which is named *average accuracy*.
2. Using the so-called *joint accuracy* (Rodríguez and Lozano, 2010) proposed in van der Gaag and de Waal (2006), where a case is classified correctly if all the class variables are labelled correctly simultaneously.

The *Brier score* can be generalized to be considered in two variants:

1. *Average Brier score*:

$$\frac{1}{Nm}\sum_{k=1}^{N}\sum_{j=1}^{m}\sum_{l=1}^{r_j}(p_{jl}^k - y_{jl}^k)^2$$

   where m is the number of class variables, r_j is the number of values of the j^{th} class variable and, similarly to the uni-dimensional *Brier score*, p_{jl}^k is the estimated probability of the single class variable C_j takes its r_j value given the k^{th} case, i.e. $p(C_j = c_{jl}|\mathbf{x}^k)$; finally, the y_{jl}^k value is 1 if l is the class observed (correct) value and 0 otherwise.

2. *Joint Brier score*:

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{g=1}^{r_o}(p_g^k - y_g^k)^2$$

   where $r_o = r_1 \times \cdots \times r_m$ is the Cartesian product of class variables values, p_g^k is the estimated probability of the g^{th} class combination for the k^{th} case, i.e. $p(\mathbf{C} = \mathbf{c}_g|\mathbf{x}^k)$; and the y_g^k value is 1 if the observed (correct) values of the m class variables correspond with the g^{th} vector of class combination and 0 otherwise.

Average Brier score and *joint Brier score* are two generalizations of *Brier score* for the multi-dimensional approach, which give different and useful information. Both consider the estimated probability assigned to the classes.

However, the *average Brier score* rewards the observed classes separately; whereas the *joint Brier score* rewards only the estimated probability of being right in all the classes simultaneously. However, neither of the generalizations rewards the number of observed values (correctly labeled classes) of each class combination, i.e. the score should be lower (superior behaviour) when higher probabilities are assigned to class variables combinations that contain two observed values, than if it contains only one observed value.

Therefore, a new score that shows this behaviour is proposed. This has been named multi-dimensional *calibrated Brier score* (MdCBS). It ranges between 0 and 1, where the lower the value the better the classifier:

$$MdCBS = \frac{1}{N * r_s} \sum_{k=1}^{N} \sum_{g=1}^{r_o} p_g^k * f_g^k$$

where p_g^k is the estimated probability of the g^{th} class combination and f_g^k is the number of non-observed class values (failures) within the current (g^{th}) combination of class variables being evaluated. This score is used in the evaluation of multi-dimensional results for the real domain dataset in Section 6.5.

The MdCBS is not used in the experimentation with artificial domains for the uni-dimensional and multi-dimensional approaches comparison, since it cannot be computed in the uni-dimensional approach. The *Joint Brier score* is also not used due to the exposed limitation.

6.3 Pre-processing methods for multi-dimensional classification

There are two basic approaches that allow the extension of uni-dimensional pre-processing methods to the multi-dimensional approach: 1) applying the uni-dimensional method to each class variable separately, and then combining the obtained results and 2) using the Cartesian product of all the classes, as a single class variable, to perform the pre-processing steps.

In addition to these approximations, the formulation of the method itself can be adapted in order to consider the nature of multiple class variables, which is one of the main contributions of this work. In the following sections the proposed multi-dimensional pre-processing methods are introduced.

6.3.1 Missing data imputation proposals for the multi-dimensional approach

In the case of missing data imputation, the *CMean* (CM) has been selected to be adapted for the multi-dimensional approach. Given a missing value in an instance of a feature, this method fills it with its mean (continuous variables) considering only the instances that present the same class variable label than

the instance with the missing value. In this work, all the features considered are continuous; therefore, the following explanation is limited to this case.

The proposed approaches for addressing the issue of missing data in the multi-dimensional approach are summarized below.

1. No imputation (NI); since some classification models can be learnt in the presence of missing data.
2. Merge of single uni-dimensional class imputations; by means of averaging the resulting imputed value for each class variable separately (CMindiv).
3. Imputation targeting the Cartesian product of classes (CMcart).

6.3.2 Discretization for the multi-dimensional approach

The supervised discretization of features is the transformation of continuous variables into categorical variables, taking into account the class values. The Fayyad and Irani's Multi-Interval Discretization (MID) method (Fayyad and Irani, 1993) has been selected. This method searches recursively, in each feature, for a set of cut-off points that reduce the class entropy. This method firstly searches for the cut-off point of the given feature X_i that minimizes the conditional entropy $H(C|X_i)$ of the class variable C. In following recursive searches, the method repeats the process on both sides of the previous selected cut-off point. The process is stopped if the gain in entropy reduction $H(C) - H(C|X_i)$ is below a Minimum Description Length (MDL) criterion (Rissanen, 1978):

$$gain > \frac{1}{N}\left(\log_2(N-1) + \Delta\right)$$

$$\Delta = \log_2(3^r - 2) - [rH(S) - r^1 H(S_{left}) - r^2 H(S_{right})]$$

where r is the number of class values present in the full training data S and r_1, r_2 are the number of class values in each resultant data subset after applying a cut-off point (S_{left}, S_{right}).

This work proposes to adapt the uni-dimensional MID to the multi-dimensional scenario as listed below.

1. The merging of single uni-dimensional class discretizations (MIDindiv).
2. Discretization targeting a single class variable formed by the Cartesian product of all class variables (MIDcart).
3. A discretization policy which considers the mean of class entropies in the MDL criterion (MIDmean), where the mean of class entropies and the mean of conditioned class entropies are used to evaluate the entropy reduction before and after adding a cut-off point (m denotes number of class variables):

$$\Delta = \log_2(3^{r_s} - 2) - \frac{1}{m}[r_s \sum_{j=1}^{m} H(S) - r_s^1 \sum_{j=1}^{m} H(S_{left}) - r_s^2 \sum_{j=1}^{m} H(S_{right})]$$

where r_s the sum of number of class values in all the class variables in the dataset or in the subdataset (r_s^1, r_s^2).
In addition, the gain in the MDL criterion has to be m times higher to accept the cut-off point:

$$gain = \frac{1}{m}\left(\sum_{j=1}^{m} H(C_j) - \sum_{j=1}^{m} H(C_j|X_i)\right)$$

4. A discretization policy which considers the sum of class entropies in the MDL criterion (MIDsum):

$$\Delta = \log_2(3^{r_s} - 2) - [r_s \sum_{j=1}^{m} H(S) - r_s^1 \sum_{j=1}^{m} H(S_{left}) - r_s^2 \sum_{j=1}^{m} H(S_{right})]$$

where the gain in the MDL criterion has to be m times higher to accept the cut-off point as in MIDmean.

6.3.3 Feature subset selection for the multi-dimensional approach

Feature subset selection (FSS) Saeys et al. (2007); Guyon et al. (2007) is the process of reducing the number of features before learning a classifier. The popular multivariate Correlation-based Feature subset Selection (CFS) method Hall (2000) has been selected in this work as a prior step to classifier learning. CFS is based upon an interesting formulation, the assumption that a good subset of forecasting features is one that is highly correlated with the class and, at the same time, the features have low correlation between them. CFS gives a merit to each feature subset $(X_{i_1}, ..., X_{i_z})$, where the correlation of each feature (in the subset) with the class is viewed positively (numerator), whilst correlation between pairs of features (in the subset) is penalised (denominator):

$$Merit(X_{i_1}, ..., X_{i_z}) = \frac{z \cdot t_{CX}}{\sqrt{z + z(z-1)t_{XX}}}$$

where $\{X_{i_1}, ..., X_{i_z}\} \subseteq \{X_1, ..., X_n\}$ being z the number of features in the subset, t_{CX} the average class-feature correlation and t_{XX} the average feature-feature correlation of a feature subset. Correlation between two variables is calculated by means of the previous exposed *symmetrical uncertainty score* (Hall, 1999).

This work proposes to adapt the uni-dimensional CFS to the multi-dimensional scenario as summarized below.

1. To select the union of all variable subsets that have been selected by the uni-dimensional CFS for each class variable in isolation (CFSindiv).

2. The selection of features targeting the Cartesian product of the m class variables (CFScart).
3. The merit of the CFS formulation is modified by considering in the numerator the mean of correlations between each class variable and each feature of the subset (CFSmean):

$$Merit(X_{i_1}, ..., X_{i_z}) = \frac{\frac{z}{m} \sum_{j=1}^{m} t_{C_j X}}{\sqrt{z + z(z-1) t_{XX}}}$$

4. The merit of the CFS formulation is modified by considering in the numerator the sum of correlations between each class variable and each feature of the subset (CFSsum):

$$Merit(X_{i_1}, ..., X_{i_z}) = \frac{z \sum_{j=1}^{m} t_{C_j X}}{\sqrt{z + z(z-1) t_{XX}}}$$

6.4 Experiments with synthetic data

The purpose of the experiments described in this section is to empirically test the behaviour of the different proposed multi-dimensional pre-processing strategies by themselves, and their joint behaviour in a pipeline. The experimentation is performed for a broad range of datasets with different intrinsic data characteristics, in order to know which strategies perform the best and under which conditions.

This is accomplished by means of statistical tests and a process of meta-learning (Hall, 1999; Inza et al., 1999; Witten and Frank, 2005). In order to proceed with this, a schema to generate synthetic data for multi-dimensional domains and procedures for methods comparison are described.

6.4.1 Synthetic data generation schema

Most experiments with continuous synthetic data are restricted to certain distribution assumptions (Hall, 1999). However, real-world problems do not necessarily follow a parametric distribution. Besides, the underlying distribution is usually unknown in real domains. There is no way of generating all possible kinds of distributions. However, kernel-based *Bayesian networks* (KBNs) can represent a broad range of distributions due to the flexibility of the kernel-based density functions (Pérez et al., 2009). The generation of synthetic domains, based upon KBNs, ensures a wide range of density shapes and different intrinsic data characteristics. This permits to reach conclusions about the proposed methods that can be considered general enough (Pérez et al., 2009).

In this framework, each domain generated is specified by the 'a priori' distribution of the discrete class variables $p(C_1), \cdots, p(C_m)$ and the density

function $\rho(X_i|\mathbf{Pa}_i)$ of each (continuous) feature X_i, given its parents $\mathbf{Pa}_i$. These densities are modelled using Gaussian kernel-based functions (Wand and Jones, 1995; Silverman, 1986). The kernel-based functions used depend upon the number of kernels q, and on the coordinates of each of those kernels $\{e_1, \cdots, e_q\}$. They also depend on a unique smooth parameter h (in this work $h = q^{-\frac{1}{6}}$, based upon Pérez et al. (2006)) that is used to define the covariance matrix $H = h^2 * \mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. The kernel-based density (Pérez et al., 2009) in its most general form can be written as:

$$f(\mathbf{x}) = \frac{1}{|q|} \sum_{k=1}^{|q|} K_H(\mathbf{x} - \mathbf{e}_k)$$

where K_H is a Gaussian kernel, defined as:

$$K_H(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |H|^{-\frac{1}{2}} exp(-\frac{1}{2}\mathbf{x}^T H^{-1}\mathbf{x})$$

Based upon the procedure presented in Pérez et al. (2006), the methodology used for the generation of synthetic datasets consists of the following steps:

1. Generate a set of KBNs input parameter values at random (Fig. 6.1).
2. Build a classifier based on a KBN, with these parameter values (Fig. 6.1).
3. Sample the generated KBN to form a dataset with a randomly-selected number of cases in {25, 50, 100}.
4. Missing values are generated in the features (class variables are excluded). The degree of missing values generation is a randomly-selected percentage {0%, 10%, 25%}.

The number of generated datasets is 1000.

The experimental set-up has been restricted to the range of parameters (Fig. 6.1) that are common in fish recruitment datasets. As it has been pointed out before, a meta-learning process is carried out to know under which conditions each pipeline and method performs the best. In order to accomplish that, for each dataset, a set of intrinsic data characteristics (e.g. missing data level or class entropy) are measured. Several of them are based upon Information Theory (see Appendix). This process allows to select the methods to be applied to real-world datasets that share the same intrinsic data characteristics.

6.4.2 Procedures for methods comparison

In order to understand which multi-dimensional pre-processing methods are superior and under which intrinsic data characteristics, several comparison scenarios are considered (Table 6.1). First, full multi-dimensional pipelines of the proposed methods are compared between each other in each pre-processing step ($COMP_{mul-pipe}$). Secondly, the uni-dimensional and the

Inputs
The number of class variables m in $\{2, 3, 4\}$, number of class values r_j in $\{2, 3, 4\}$, number of features n in $\{5, 10, ..., 25\}$, maximum number of feature parents for each feature in $\{1, 2\}$, probability of an arc between a class and a feature $p(C_j \rightarrow X_i)$, probability of an arc between class variable pairs $p(C_j \rightarrow C_{j+1})$.
Outputs
A probabilistic model M based on KBN.
Algorithm
Generate randomly a graph G of the KBN structure taking into account the input parameters.
For each feature $X_i \in \{X_1, ..., X_n\}$:
For each configuration of the discrete parents, $\mathbf{pd}_i$, of X_i:
A number of kernels $q_{\mathbf{pd}_i}$ is randomly selected from $\{1, 2, 4, ..., 512, 1024\}$.
A set of kernels $q_{\mathbf{pd}_i}$ is built, sampling $q_{\mathbf{pd}_i}$ points $(x_i, \mathbf{pc}_i)$ from a uniform distribution in [1, 10].
End for
End for

Fig. 6.1. The pseudocode for the generation of a classifier, based upon a KBN. $\mathbf{pd}_i$ denotes the discrete parents of X_i and $\mathbf{pc}_i$ is used to denote its continuous parents, and where $\mathbf{pa}_i = \mathbf{pd}_i \cup \mathbf{pc}_i$.

multi-dimensional approaches are compared ($COMP_{uni-mul}$), in particular the best multi-dimensional pipelines are compared with the uni-dimensional pipeline. Finally, the proposed pre-processing methods are compared with each other ($COMP_{mul-pre}$): missing data imputation ($COMP_{mul-mis}$), discretization ($COMP_{mul-dis}$) and feature subset selection ($COMP_{mul-FSS}$).

Table 6.1. General overview of methods comparison schema. In first column the two main levels of comparison ($COMPove$ and $COMPspe$), in the second the three scenarios ($COMP_{mul-pipe}$, $COMP_{uni-mul}$ and $COMP_{mul-pre}$) and in the third column the three measures used for these comparisons ($COMP^{avAcc}$, $COMP^{jAcc}$ and $COMP^{avBS}$).

Comparison level	Comparison scenarios	Measures of performance
$COMPove$: Identification of pipelines, approaches (Uni-D, Mul-D) and pre-processing methods, that show an overall superior behaviour	$COMPove_{mul-pipe}$: Multi-dimensional pipelines	^{avAcc}: *average accuracy* ^{jAcc}: *joint accuracy* ^{avBS}: *average Brier score*
	$COMPove_{uni-mul}$: Uni-D *vs* Mul-D approaches	^{avAcc}, ^{jAcc} and ^{avBS} (superindex)
	$COMPove_{mul-pre}$: Mul-D pre-processing methods (subindex): $_{mul-mis}$, $_{mul-dis}$ and $_{mul-FSS}$	^{avAcc}, ^{jAcc} and ^{avBS} (superindex)
$COMPspe$: Identification of intrinsic data characteristics where each pipeline, method or approach (Uni-D, Mul-D) have specific superior behaviour	$COMPspe_{mul-pipe}$	^{avAcc}, ^{jAcc} and ^{avBS} (superindex)
	$COMPspe_{uni-mul}$	^{avAcc}, ^{jAcc} and ^{avBS} (superindex)
	$COMPspe_{mul-pre}$ (subindex): $_{mul-mis}$, $_{mul-dis}$ and $_{mul-FSS}$	^{avAcc}, ^{jAcc} and ^{avBS} (superindex)

The comparison is performed at two different levels of detail (Table 6.1, first column). On the one hand, it is established which pipelines, approaches and methods show a superior overall behaviour ($COMPove$). On the other, a meta-learning process is performed to determine under which intrinsic data

characteristics (Fig. 6.1 and Appendix) each method shows a specific superior behaviour ($COMPspe$). These two levels of comparison are performed for the three scenarios previously presented.

In addition, the comparison is performed for three multi-dimensional performance measures, which are denoted by the following superindex (Table 6.1, third column):

- ^{avAcc}; for comparison in terms of *average accuracy*;
- ^{jAcc}; for comparison in terms of *joint accuracy*;
- ^{avBS}; for comparison in terms of *average Brier score*.

The overall comparison at the first level ($COMPove$) is performed by means of the revised Friedman plus Shaffer's static post-hoc test, proposed by García and Herrera (2008) for comparison of multiple methods over multiple datasets. These statistical test results can be represented by means of critical difference diagrams Demŝar (2006), which show the average ranks of the performance of each method across all the domains in a numbered line. If there is not a statistically significant difference between two methods, they are connected in the diagram by a straight line. As an example, in Figure 6.5, NI and CMindiv methods are connected since they show no significant difference for *average accuracy*; whereas these methods are unconnected for *average Brier score*, showing a statistically significant difference.

The specific data characteristics comparison at the second level ($COMPspe$) is performed by studying the intrinsic data characteristics of the sampled datasets as forecasting factors (Fig. 6.1 and Appendix) in a meta-learning process (Witten and Frank, 2005; Inza et al., 1999). For this purpose, a meta-dataset (Table 6.2) is compiled with the intrinsic data characteristics of all the simulated synthetic datasets. For each of the previously mentioned comparison scenarios a target node is added to the meta-dataset with the result of these comparisons. For example, in the $COMPspe^{avAcc}_{mul-pipe}$ scenario, a target variable is added such that for the i^{th} dataset it takes the value of the best pipeline (Table 6.2). As a way of studying this dataset, the 'Markov blanket'[1] of the target node of this meta-dataset is of interest in order to determine under which characteristics a method is superior to others. Each 'Markov blanket' is learned using the method proposed in Peña et al. (2007). Instead of representing each 'Markov blanket' of each target variable in a different figure, all the 'Markov blanket' are jointly represented in a single structure for each scenario comparison (Fig. 6.3, 6.4 and 6.6) as a way of summarising the results.

[1] In a Bayesian network, the 'Markov blanket' of a node includes its parents, children and the other parents of all of its children **?**. Therefore, the 'Markov blanket' of a variable (X) is the smallest set ($Mb(X)$) containing all variables carrying information about X that cannot be obtained from any other variable Peña et al. (2007); Pellet and Elisseeff (2008), meaning that the variables of the 'Markov blanket' is the only knowledge base needed to forecast the behaviour of the target variable.

The 'Markov blanket' is used to perform inference and extract useful information; i.e., the value of one or several variables (both; characteristics or target nodes) can be fixed and the effects on the rest observed. For example, under which intrinsic data characteristics a method (e.g. MIDmean) shows a superior behaviour can be determined; or given a value of an intrinsic data characteristic (e.g. evidence of 10% missing values), the behaviour of the different methods observed. Results that accomplished certain probability thresholds have been considered useful knowledge. In the case of target variables (for each performance measure) with three or four values (for each method that shows a superior behaviour), evidence needs to reach at least 0.5 probability. Whereas in the case of pipelines evaluation, the target nodes (one for each performance measure) have 48 values (one for each possible pipeline), and only results with a probability higher than 0.1 have been used.

Table 6.2. Example of the meta-dataset for meta-learning of characteristics related with the superiority of a specific pipeline/approach/method over the rest. The dataset is composed of intrinsic data characteristics of the sampled datasets and a set of target nodes, whose values contains the pipeline/approach/method that has a higher performance for each dataset.

	Intrinsic data characteristics						Target nodes of the meta-analysis ($COMPspe$)		
Set	# cases	Miss. level	...	Class entr.	Class inter.	Feat. relev.	...	$COMP^{avAcc}_{mul-pipe}$ ...	$COMP^{avBS}_{pre-FSS}$ ...	$COMP^{avBS}_{uni-mul}$
1	50	10		0.61	0.21	0.57		CMcart-MIDcart-CFSsum	CFSsum	Uni-D
2	100	10		0.57	0.06	0.45		NI-MIDsum-CFSindiv	CFSindiv	Uni-D
3	50	0		0.64	0.15	0.65		CMcart-MIDmean-CFScart	CFScart	Mul-D
4	25	25		0.43	0.32	0.32		CMindiv-MIDsum-CFScart	CFScart	Uni-D
5	100	0		0.69	0.09	0.73		CMcart-MIDmean-CFSmean	CFSmean	Mul-D
...	...	...		...	...	...		...	...	...
1000	50	25		0.54	0.17	0.24		CMcart-MIDsum-CFSsum	CFSsum	Uni-D

6.4.3 Software

All of the above steps have been implemented using several established API machine-learning software tools: Weka (Witten and Frank, 2005) has been adapted to implement the MDnB classifier and the multi-dimensional preprocessing methods, as well as to calculate the performance measures considering multiple class variables; and the Elvira software platform (Consortium, 2002), using Jama matrix library, which was adapted in a previous work to generate synthetic data based on *flexible Bayesian network* classifiers in Pérez et al. (2009), has been readapted to generate synthetic data with multiple class variables. Reproducibility is ensured by a Java programming language implementation of all the methodology using these APIs. This software is available from the ISG group and Azti-Tecnalia webpages (*www.sc.ehu.es/ccwbayes/members/jafernandes* or *www.azti.es*).

6.4.4 Results on synthetic data

In the following sections, the results on synthetic data are presented for the three comparison scenarios (Table 6.1): pipelines of multi-dimensional methods ($COMP_{mul-pipe}$), uni-dimensional *vs* multi-dimensional approaches ($COMP_{uni-mul}$) and the multi-dimensional pre-processing methods in each pipeline step ($COMP_{mul-pre}$). The comparison is firstly performed observing which methods show an overall superior behaviour ($COMP_{ove}$) and secondly identifying under which conditions each method outperforms the others ($COMP_{spe}$).

6.4.4.1 Comparison of multi-dimensional pipelines based on proposed methods

In this section, the results of multi-dimensional pipelines of methods are analysed ($COMP_{mul-pipe}$). A pipeline is considered to be the combination of a missing data imputation method, a discretization technique, a feature subset selection strategy and a MDnB classifier. Although the large number of method combinations make it difficult to extract conclusions, some general trends can be observed.

Figure 6.2 shows the results of overall behaviour comparison for pipelines of multi-dimensional methods ($COMPove_{mul-pipe}$). The figure represents the critical diagram obtained with the 1000 simulated synthetic datasets. Pipelines that contain the CMCart imputation and MIDmean or MIDsum discretization methods tend to have a superior behaviour than the rest.

In the case of specific data characteristics comparison ($COMPspe_{mul-pipe}$; Fig. 6.3), such as high levels of missing values, it can be observed that pipelines with MIDmean show a superior behaviour than the rest of Mul-D discretization methods in terms of both *average accuracy* and *joint accuracy*. In datasets where there are no missing values, the use of MIDmean method shows a superior behaviour. The no imputation, in combination with MIDcart and MIDmean, is another pipeline which shows a solid behaviour. Finally, in terms of *average Brier score* ($COMPspe^{avBS}_{mul-pipe}$), the pipelines with better behaviour are those that include CMindiv-MIDmean with CFSsum or CFSindiv feature selection methods in the case of high levels of missing data; whereas with moderate levels of missing data, the pipelines that contain CMcart are the ones that show superiority over the others.

6.4.4.2 Comparison of the uni-dimensional vs the multi-dimensional approaches

In the case of overall comparison ($COMPove_{uni-mul}$), the uni-dimensional (Uni-D) pipeline statistically outperforms the multi-dimensional (Mul-D) approach (Fig. 6.2) in terms of accuracy measures (*average* and *joint*). However, the difference in ranks between both approaches is lower in *joint accuracy*

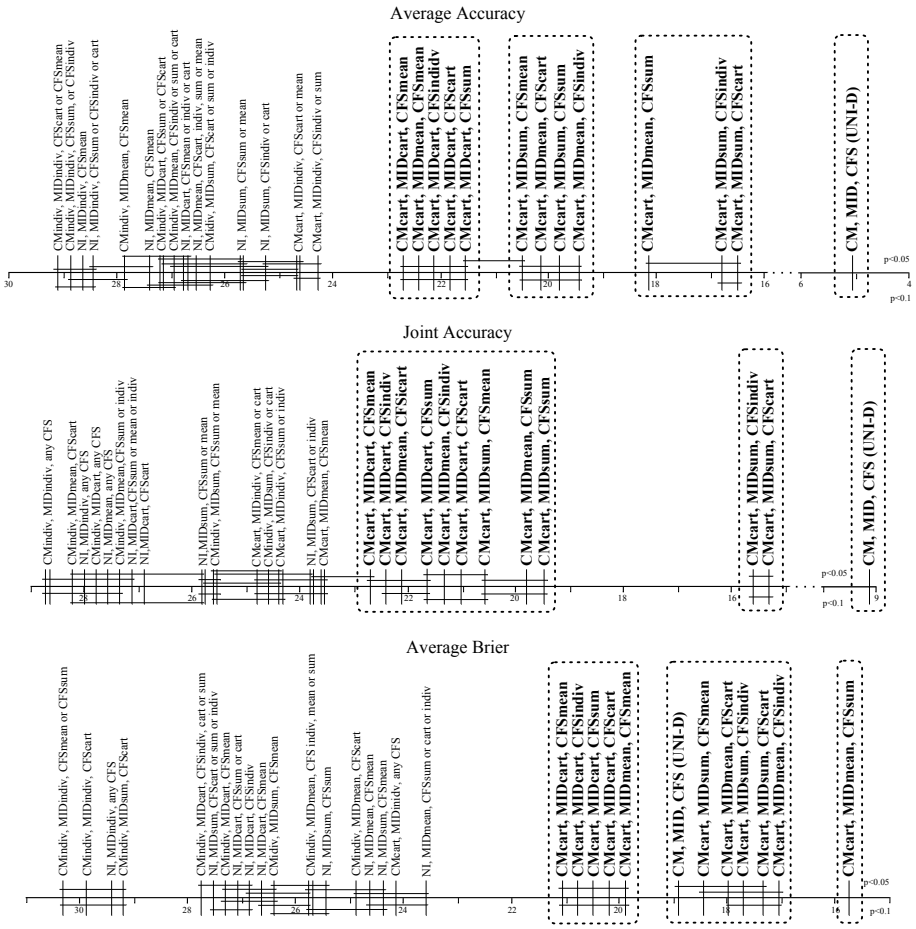


Fig. 6.2. Critical difference diagrams, comparing pipelines of multi-dimensional methods and the uni-dimensional pipeline in synthetic datasets for the three performance measures. The methods with a general superior behaviour are shown on the right hand side of each diagram. Pipelines that do not show a statistical significant difference are connected by a horizontal line.

than in *average accuracy*; this means that Mul-D pipelines have a superior behaviour in terms of *joint accuracy* for many datasets.

In terms of *average Brier score* ($COMPove^{avBS}_{uni-mul}$), several of the Mul-D pipelines outperform the Uni-D approach, as can be observed in the right side of the third critical difference diagram, in Figure 6.2. In particular, the Mul-D pipelines that contain the CMcart missing data imputation strategy show a superior behaviour. The fact that the multi-dimensional approach shows a superior behaviour in *average Brier score* is of importance in domains such as recruitment forecast (Fernandes et al., 2010c), where the value of the 'a

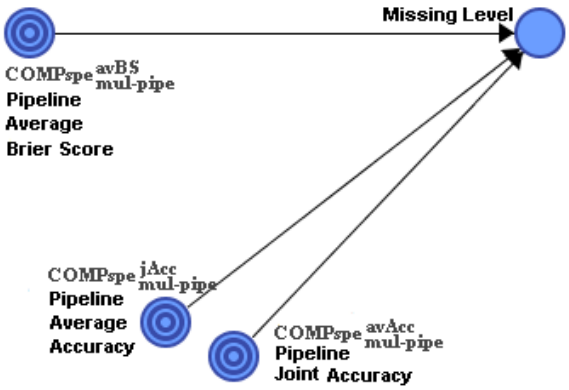


Fig. 6.3. Results of the meta-learning process represented in a single structure for multi-dimensional pipelines ('$COMPspe_{mul-pipe}$'). A single characteristic (level of missing data) forms the 'Markov Blanket' of the three target nodes, determining the superior behaviour of some pipelines in comparison with the rest in terms of *average accuracy*, *joint accuracy* and *average Brier score*.

posteriori' probabilities for each class value is crucial for informed management decision-making.

Missing data level is the specific characteristic ($COMPspe_{uni-mul}$) that stands out as the most influential for the superiority of an approach when *average accuracy* is used as score. This is also a key characteristic in terms of *joint accuracy*, which is influenced also by the class entropy and the total number of class values.

In terms of *average Brier score*, the most influential characteristics are the class entropy and the total number of class values. These specific conditions, where the Mul-D approach is superior to the Uni-D scheme, can be analysed using the learned 'Markov blanket' set (Fig. 6.4), fixing the Mul-D value of the target nodes and performing inference:

- In terms of *average accuracy* ($COMPspe^{avAcc}_{uni-mul}$); when the total number of class values is low.
- In terms of *joint accuracy* ($COMPspe^{jAcc}_{uni-mul}$); when the total number of class values as well as the class entropy are low and the missing data level is high.
- In terms of *average Brier score* ($COMPspe^{avBS}_{uni-mul}$); when the total number of class values is high and with low levels of class entropy.

6.4.4.3 Comparison of multi-dimensional missing imputation methods

In terms of missing data imputation strategies, observing critical difference diagram results (Fig. 6.5), CMcart is, in general, superior ($COMPove_{mul-mis}$) in all of the multi-dimensional performance measures.

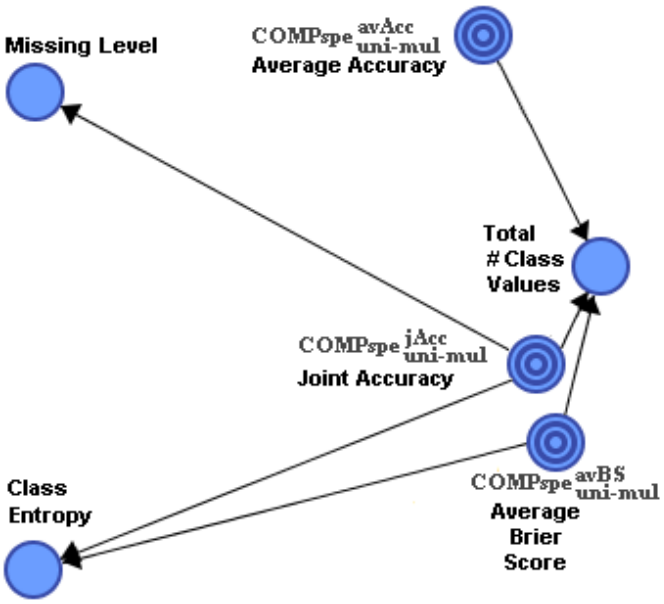


Fig. 6.4. Results of the meta-learning process, summarised in a single structure, for the comparison of specific conditions ('$COMPspe_{uni-mul}$'). It shows whether the uni-dimensional or the multi-dimensional pipeline is superior, for the three multi-dimensional performance measures. As an example, '$COMPspe^{avBS}_{uni-mul}$' node values are the approaches that show a superior behaviour (Uni-D or Mul-D) in *average Brier Score* for each generated synthetic dataset; its 'Markov blanket' is composed of two nodes. The target nodes of the two learned 'Markov blanket' are emphasised.

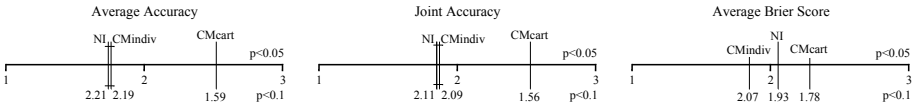


Fig. 6.5. Critical difference diagrams comparing multi-dimensional missing data imputation methods in synthetic datasets for three performance measures. Methods that do not show a significant difference are connected in the diagram.

However, there are specific conditions (Fig. 6.6) where each method shows a superior behaviour ($COMPspe_{mul-mis}$). This can be observed fixing each target node value and performing inference over the characteristics in its 'Markov blanket'.

In terms of *average accuracy* ($COMPspe^{avAcc}_{mul-mis}$) and *joint accuracy* ($COMPspe^{jAcc}_{mul-mis}$):

- CMcart, with higher levels of missing values.
- CMindiv, with lower levels of missing values.

In terms of *average Brier score* ($COMPspe^{avBS}_{mul-mis}$):

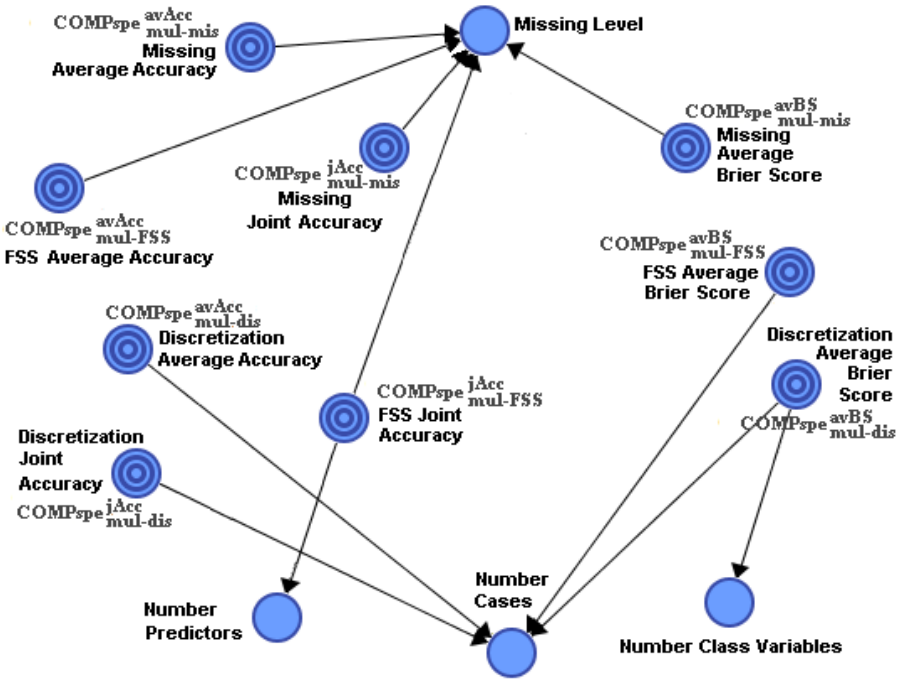


Fig. 6.6. Meta-learning process results, represented in a single structure, show that only four characteristics (plain nodes) are the most influential in terms of *average accuracy*, *joint accuracy* and *average Brier score*, for all multi-dimensional pre-processing steps (emphasized nodes). As an example, 'FSS average BS' node contains the feature subset selection method that shows a superior behaviour in *average Brier Score*, for each generated synthetic dataset; its 'Markov blanket' is composed of a single node (Number of Cases).

- CMcart, with lower levels of missing values.
- CMindiv, with higher levels of missing values.

6.4.4.4 Comparison of the proposed multi-dimensional discretization methods

Based upon critical difference diagram results (Fig. 6.7), overall behaviour of the discretization methods can be observed ($COMPove_{mul-dis}$). The MIDmean method shows the best behaviour for *average accuracy* and *joint accuracy*, followed by MIDsum, MIDcart and MIDindiv. In the case of *average Brier score*, MIDmean shows the worst behaviour.

However, there are specific conditions (Fig. 6.6) where each method shows a superior behaviour ($COMPspe_{mul-dis}$).

In terms of accuracy measures ($COMPspe^{avAcc}_{mul-dis}$ and $COMPspe^{jAcc}_{mul-dis}$):

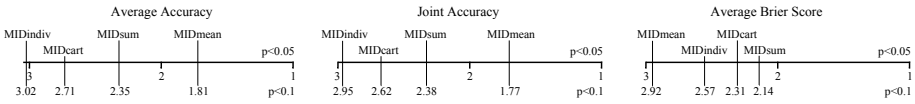


Fig. 6.7. Critical difference diagrams, comparing multi-dimensional discretization methods in synthetic datasets for three performance measures. All the methods show a statistically significant difference each other; therefore, they are not connected in the diagram.

- MIDsum and MIDcart, with a low number of cases.
- MIDindiv shows a superior behaviour in datasets with few cases.

In terms of *average Brier score* ($COMPspe^{avBS}_{mul-dis}$):

- MIDmean and MIDcart, with high number of class variables and with few cases.
- MIDsum, with few class variables and high number of cases.

6.4.4.5 Comparison of proposed multi-dimensional feature selection methods

The CFScart method shows the best overall behaviour ($COMPove_{mul-FSS}$) in terms of *joint accuracy* based on the critical difference diagram (Fig. 6.8). However, it shows a tie with CFSsum in *average accuracy* terms. CFSmean is the best contender in terms of *average Brier score*. In terms of specific behaviour, each method shows superiority to the rest ($COMPspe_{mul-FSS}$) under different intrinsic data characteristics (Fig. 6.6):

- CFSmean, with low levels of missing values, few features and with a high number of cases.
- CFSsum, with high levels of missing values and a moderate amount of cases.
- CFSindiv, with few cases in terms of *Brier score* and moderate levels of missing values in terms of *joint accuracy*.
- CFScart, with high levels of missing values, a high number of features and a moderate amount of cases.

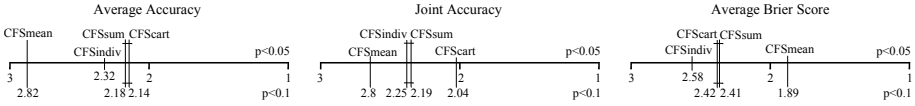


Fig. 6.8. Critical difference diagrams, comparing multi-dimensional feature subset selection methods in synthetic datasets. Methods that do not show a significant difference are connected by a horizontal line.

6.5 Application to fish recruitment forecasting

In this section, the application of the multi-dimensional approach to a real domain of multi-dimensional nature is presented. This domain is fish recruitment forecast (Fernandes et al., 2010c), where recruitment is the amount of fish that enters the fishery each year. Recruitment is considered more important than the total stock because it represents the future of the fish population, conditioning the feasibility of species exploitation (Fernandes et al., 2010c).

This is a domain of high *uncertainty* and with considerable biological, social, economic and political impact. In this study, a subset of the proposed methods is tested with three species that are of commercial interest in the Bay of Biscay; anchovy, sardine and hake. The nature of this domain, with interactions due to the sharing ecosystem, demands a multi-dimensional approach.

In order to apply the multi-dimensional approach to species of the Bay of Biscay, three scenarios are considered since the relationships between the species have not been established by experts. In a first multi-dimensional study, two class variables are considered; anchovy recruitment index (ARI) and hake recruitment (HR) (Table 6.3). Anchovy is a short-life fish, likely to be a prey of hake, which is a long-life species. In a second multi-dimensional study, another pair of class variables are considered: ARI and SR (sardine recruitment); sardine is suspected to be a competitor of anchovy. Finally, three class variables are considered: ARI, HR and SR, where hake might be a predator of anchovy and sardine, whereas sardine might be a predator of the two other species eggs.

Table 6.3. Intrinsic data characteristics of fish recruitment datasets, for the three species recruitment: anchovy (ARI), sardine (SR) and hake (HR).

Characteristic	ARI-HR	ARI-SR	ARI-SR-HR
# cases	41	41	41
Missing level (%)	0.2	0.2	0.2
# class variables	2	2	3
Total # class values	9	9	27
# features	192	192	192
Class entropy	0.67	0.65	0.66
Classes interaction	0.06	0.07	0.14

Anchovy and hake are species of high commercial interest in the Bay of Biscay that share the ecosystem with sardine and, consequently, they share the same factor candidates. Fish recruitment is originally a continuous variable, which has been discretized using the method proposed in Fernandes et al. (2010c) to work with levels of recruitment; this is considered adequate for management decision-making.

The previous experiments with synthetic datasets are used to select the multi-dimensional methods to be applied in this real domain scenario (Ali and Smith, 2006) of intrinsic multi-dimensional nature. In terms of overall behaviour (Fig. 6.2), the pipeline CMcart-MIDsum-CFScart shows a superior behavior in *average accuracy* and *joint accuracy*, whereas in terms of *average Brier score*, the pipeline with superior behaviour is CMcart-MIDmean-CFSsum.

In general, pipelines that contain any of CMcart and MIDmean or MIDsum methods show an overall superior behaviour. In addition, considering the domain intrinsic data characteristics (Table 6.3), such as few cases or low class entropy, other methods can have a superior behaviour: CMindiv missing data imputation and MIDindivdiscretization. Therefore, in addition to the mentioned CMcart-MIDsum-CFScart and CMcart-MIDmean-CFSsum pipelines, the following pipelines have been selected to be used in the real domain based upon the plethora of results:

- CMcart-MIDindiv-CFScart
- CMcart-MIDmean-CFSmean
- CMcart-MIDmean-CFSindiv
- CMcart-MIDmean-CFScart
- CMindiv-MIDindiv-CFSmean
- CMindiv-MIDindiv-CFSsum
- CMindiv-MIDindiv-CFSindiv

6.5.1 Anchovy and hake multi-dimensional modelling

In Table 6.4, the results of the application of the uni-dimensional pipeline (Uni-D) and the set of selected multi-dimensional pipelines (Mul-D) are shown. It can be observed that the CMcart-MIDmean-CFSsum combination is the one with the highest *joint accuracy*, showing a significant difference with the Uni-D approach ($p<0.05$; *corrected paired t-test*, Nadeau and Bengio (2003)). It also shows a higher *average accuracy*, in comparison to the Uni-D pipeline. In addition, this combination shows the highest accuracy for anchovy (non-significant) and the lowest *Brier score* for hake ($p<0.05$) in comparison with the Uni-D approach. However, a high number of features are selected, as well as a high number of intervals for each feature.

An interesting point is that some of the Mul-D methods superior performance could be due to selecting a higher number of features and intervals (Kohavi and Sommerfield, 1995; Kononenko, 1995). However, some of the combinations show a general tendency to overcome the Uni-D approach with the same or lower number of features and intervals (Table 6.4). Moreover, the number of features, or intervals, could be limited introducing a Minimum Description Length criteria; however, this could be a questionable solution (Domingos, 1999), in such a domain.

In addition to performance improvement, model comprehension and transparency is a key issue for domain experts in order to be able to extract knowl-

edge, or to obtain a descriptive model of the relationship between the environment and the recruitment of the species of interest. As such, an intermediate approach could be to consult the full table of results with experts, so they can decide a proper balance depending upon the research objective (Fernandes et al., 2009c, 2010c).

As an example, the combination CMcart-MIDmean-CFSsum (Table 6.4) shows a large number of selected features. However, the combination CMcart-MIDmean-CFSmean shows a similar performance, with only 22 features (instead of 86) and less intervals per feature. In addition, the pipeline CMcart-MIDmean-CFSindiv shows superior behaviour in terms of the multi-dimensional *calibrated Brier score* with only 15 features.

6.5.2 Anchovy and sardine multi-dimensional modelling

In Table 6.5, the results for the uni-dimensional pipeline (Uni-D) and multidimensional (Mul-D) pipelines, targeting anchovy and sardine species, can be observed. The pipeline CMcart-MIDmean-CFSsum shows the highest *joint accuracy* and *average accuracy* performance values, as well as a lower *average Brier score* than the Uni-D pipeline with significant statistical difference ($p<0.05$). This pipeline selects a high number of features and intervals. However, there are several pipelines that outperform the uni-dimensional approach, with a lower number of selected features (e.g. CMcart-MIDmean-CFSmean).

The pipeline with superior behaviour, in terms of the multi-dimensional *calibrated Brier score*, is the pipeline CMcart-MIDindiv-CFScart; this pipeline also shows a superior behaviour in terms of *joint accuracy*.

The CMindiv-MIDindiv-CFSindiv combination of methods shows the highest uni-dimensional *accuracy* (non-significant) and *Brier score* ($p<0.05$), with respect to the Uni-D pipeline for anchovy; whereas, in the case of sardine, the improvement is non-significant.

6.5.3 Anchovy, sardine and hake multi-dimensional modelling

The three species uni-dimensional (Uni-D) and multi-dimensional pipelines results are listed in Table 6.6. The CMcart-MIDmean-CFScart pipeline is the one with the highest *joint accuracy* ($p<0.05$), as well as better uni-dimensional *accuracy* and *Brier score* for anchovy (non-significant) in comparison with the Uni-D pipeline. This Mul-D pipeline also improves the results of the Uni-D approach for hake and sardine, in both *accuracy* (non-significant) and *Brier score* ($p<0.05$).

The combination CMcart-MIDmean-CFSsum shows the highest *average accuracy* (non-significant) and *accuracy* for hake ($p<0.05$). Both pipelines return a high number of selected features. However, the combinations CMcart-MIDindiv-CFScart and CMcart-MIDmean-CFSmean show the lowest *average*

Table 6.4. Pipelines performance results are shown for anchovy and hake recruitment study. In the first row, uni-dimensional approach pipeline results are shown, followed by multi-dimensional pipelines in the rest of rows. The uni-dimensional *Brier score* (BS) has been divided by 2 in order to be in the more comprehensible range (0-1). In the last three columns, the multi-dimensional *calibrated Brier score* (MdCBS), the number of features (Fea.) and average discretization intervals (Int.) selected are shown. The best results are emphasised in bold.

Pre-processing pipeline	ARI Acc.	ARI BS	HR Acc.	HR BS	Acc. Av.	Joint Acc.	BS Av.	MdCBS	Fea.	Int.
CM-MID-CFS (Uni-D)	52.7 ± 6.7	0.36	67.6 ± 3.3	0.27	60.2 ± 10.5	30.4 ± 5.9	0.32	–	16	2.1
CMcart-MIDsum-CFScart	50.9 ± 4	0.31	65.1 ± 5.8	0.25	58 ± 10	34.4 ± 7.4	0.28	0.57	24	10.4
CMcart-MIDmean-CFSsum	$\mathbf{55.9 \pm 5.9}$	0.34	$\mathbf{75.6 \pm 3.6}$	**0.18**	$\mathbf{65.7 \pm 14}$	$\mathbf{41.1 \pm 4.7}$	0.26	**0.39**	86	5.8
CMcart-MIDindiv-CFScart	50.8 ± 3.8	0.32	75.1 ± 8.3	**0.18**	62.9 ± 17.3	37.6 ± 5.5	**0.25**	0.48	15	2.3
CMcart-MIDmean-CFSmean	53.9 ± 5.7	0.3	74.2 ± 5.3	0.2	64.1 ± 14.3	39.1 ± 6.5	**0.25**	0.5	22	5
CMcart-MIDmean-CFScart	53.4 ± 7.7	0.32	75.2 ± 4.3	**0.18**	64.3 ± 15.4	40.1 ± 6.8	**0.25**	0.44	51	5.6
CMcart-MIDmean-CFSindiv	46.6 ± 6.8	0.35	71.2 ± 4.6	0.19	58.9 ± 17.3	33.2 ± 4.6	0.27	0.48	29	4.8
CMindiv-MIDindiv-CFSmean	54.5 ± 7.1	**0.29**	62.7 ± 6.5	0.28	58.6 ± 5.8	31.7 ± 4	0.28	0.53	15	2.1
CMindiv-MIDindiv-CFSsum	54.6 ± 6.7	**0.29**	61.9 ± 6.7	0.28	58.3 ± 5.2	31.9 ± 4	0.28	0.53	15	2.1
CMindiv-MIDindiv-CFSindiv	50.3 ± 6.6	0.32	69.1 ± 11.2	0.26	59.7 ± 13.3	31.7 ± 8.8	0.29	0.51	20	2

Table 6.5. Pipelines performance results are shown for anchovy and sardine recruitment study. In the first row, uni-dimensional approach pipeline results are shown, followed by multi-dimensional pipelines in the rest of rows. The uni-dimensional *Brier score* (BS) has been divided by 2 to be in the more comprehensible range (0-1). In the last three columns, the multi-dimensional *calibrated Brier score* (MdCBS), the number of features (Fea.) and average discretization intervals (Int.) selected are shown. The best results are emphasized in bold.

Pre-processing pipeline	ARI Acc.	ARI BS	SR Acc.	SR BS	Acc. Av.	Joint Acc.	BS Av.	MdCBS	Fea.	Int.
CM-MID-CFS (Uni-D)	52.7 ± 6.7	0.36	55.7 ± 6.7	0.34	54.2 ± 2.1	26.3 ± 8.2	0.35	–	13	2.1
CMcart-MIDsum-CFScart	48.7 ± 8.2	0.31	58.1 ± 4.9	0.26	53.4 ± 6.7	29.3 ± 5.3	0.29	0.57	20	9.4
CMcart-MIDmean-CFSsum	57.6 ± 8.8	0.31	$\mathbf{65.6 \pm 5.2}$	0.27	$\mathbf{61.6 \pm 5.6}$	$\mathbf{42.4 \pm 9}$	0.29	**0.42**	91	5.7
CMcart-MIDindiv-CFScart	51.7 ± 6.6	0.31	59.3 ± 4.5	0.26	55.5 ± 5.4	32.1 ± 5.9	0.29	0.53	15	2.1
CMcart-MIDmean-CFSmean	54.3 ± 6.5	0.29	62.6 ± 6.3	**0.25**	58.5 ± 5.9	37.1 ± 7	**0.27**	0.5	25	5.2
CMcart-MIDmean-CFScart	57.6 ± 7	0.3	61.9 ± 5.1	0.27	59.7 ± 3.1	38.7 ± 8.2	0.28	0.46	60	5.5
CMcart-MIDmean-CFSindiv	56.8 ± 6.4	0.29	60.2 ± 6	0.26	58.5 ± 2.4	37.4 ± 3.2	0.28	0.47	43	5.2
CMindiv-MIDindiv-CFSmean	54.5 ± 6.3	0.29	49.1 ± 5.9	0.32	51.8 ± 3.9	29.6 ± 8.3	0.3	0.57	14	2.1
CMindiv-MIDindiv-CFSsum	55.7 ± 4.6	0.28	49.4 ± 5.8	0.32	52.6 ± 4.5	30.5 ± 8.8	0.3	0.57	14	2.1
CMindiv-MIDindiv-CFSindiv	$\mathbf{58.2 \pm 5.2}$	**0.27**	50.5 ± 4.8	0.31	54.3 ± 5.74	29.6 ± 6.7	0.29	0.55	14	2.1

Brier score, with a low number of selected features. The pipeline with a superior behaviour, in terms of the multi-dimensional *calibrated Brier score*, is the pipeline CMcart-MIDmean-CFSindiv (similar to the ARI-HR dataset); this pipeline also shows a superior behaviour in terms of *joint accuracy*.

In general, the multi-dimensional approach improves the uni-dimensional *accuracy* of each species; however, the differences are often not statistically significant. A key issue in this domain is that the multi-dimensional approach improves in terms of the uni-dimensional *Brier score* of each species with differences that are statistically significant. In addition, the most important result is that there are notable improvements of the *joint accuracy*, when the multi-dimensional approach is used, i.e. the chance of being correct, at the same time, in all of the species is higher than using a uni-dimensional classifier for each species. In fact, the increase of *joint accuracy* is higher than *average accuracy*, or the uni-dimensional *accuracy* of each species. This simultaneous improvement in all species is crucial in terms of the ecosystem-based fisheries management approach.

6.6 Conclusions and recomendations for future work

To the best of authors' knowledge, this is the first study which proposes a set of supervised filter pre-processing methods for multi-dimensional classification. This study is complemented by the identification of the conditions where the multi-dimensional approach can be superior to the uni-dimensional approach. Similarly, the conditions where a pre-processing method is superior to others are identified. The most influential characteristics for the success of specific multi-dimensional pre-processing methods are the number of cases, the level of missing values and the class entropy. These conclusions suggest that the pre-processing methods can be improved by taking into account these characteristics in future works.

On the other hand, the principal objective of proposing a set of multi-dimensional pre-processing approaches, appropriate for fisheries management, is achieved. The application to a real oceanographic problem, which shows a clear multi-dimensional nature, reveals benefits in terms of improving forecasting of fish recruitment for management purposes. Firstly, improvement in the forecasting of each species is achieved (individual and average species *accuracy*). Secondly, a significant improvement in the chance of simultaneous correct forecast for all of the species (*joint accuracy*), which is a key issue for the ecosystems management approach, can be highlighted. Finally, significant improvement in the 'a posteriori' estimated class probabilities, which leads to better informed decisions, can be shown (single, *average* and *joint Brier score*). This is a key objective in knowledge-based fisheries management.

Finally, a line of future work will consist of learning descriptive, instead of forecasting, models for knowledge extraction purposes, i.e. building classifiers

Table 6.6. Comparison of multi-dimensional *vs* uni-dimensional approach for anchovy, sardine and hake study. Best results are emphasized in bold.

Pre-processing pipeline	ARI Acc.	ARI BS	SR Acc.	SR BS	HR Acc.	HR BS	Acc. Av.	Joint Acc.	BS Av.	MdCBS	Fea.	Int.
CM-MID-CFS (Uni-D)	52.7 ± 6.7	0.36	55.7 ± 6.7	0.34	67.6 ± 3.3	0.27	58.7 ± 7.9	17.3 ± 4.8	0.32	–	21	2
CMcart-MIDsum-CFScart	46.2 ± 5.8	0.32	54.7 ± 3.6	0.27	55.5 ± 6.6	0.27	52.1 ± 5.2	22.5 ± 3.5	0.28	0.59	37	13.4
CMcart-MIDmean-CFSsum	54.6 ± 7.3	0.35	$\mathbf{65.4 \pm 5}$	0.27	72.9 ± 5.5	0.21	$\mathbf{64.3 \pm 9}$	28.9 ± 4.5	0.28	**0.38**	138	4.5
CMcart-MIDindiv-CFScart	46.5 ± 4.3	0.32	59.8 ± 6.3	**0.24**	71.4 ± 5.8	0.19	59.2 ± 12.5	22.6 ± 4.3	**0.25**	0.5	21	2.5
CMcart-MIDmean-CFSmean	45 ± 7.6	0.32	58.4 ± 6	0.25	$\mathbf{75 \pm 4.6}$	**0.18**	59.4 ± 14.9	19.7 ± 5.5	**0.25**	0.49	27	4.2
CMcart-MIDmean-CFScart	$\mathbf{57.9 \pm 5}$	0.3	60.6 ± 4.8	0.27	68.9 ± 7.3	0.21	62.5 ± 5.8	$\mathbf{29.5 \pm 4}$	0.26	0.42	101	4.5
CMcart-MIDmean-CFSindiv	53.8 ± 4.8	0.32	63.4 ± 2.9	0.27	71.6 ± 6.1	**0.18**	63 ± 8.9	28.5 ± 4.7	0.26	0.41	52	4.3
CMindiv-MIDindiv-CFSmean	49.3 ± 4.8	0.31	44.9 ± 8.4	0.32	65 ± 4	0.26	53.1 ± 10.6	10 ± 4.3	0.29	0.58	15	2.2
CMindiv-MIDindiv-CFSsum	52.2 ± 5.9	**0.29**	43.4 ± 7.3	0.32	60.3 ± 4.6	0.28	52 ± 8.5	10.3 ± 2.5	0.3	0.58	17	2.2
CMindiv-MIDindiv-CFSindiv	52.9 ± 6.9	0.31	51.5 ± 7.5	0.3	67.5 ± 7.3	0.25	57.3 ± 8.8	14 ± 6.7	0.28	0.54	20	2.1

where the structure is not fixed. These models can help in the process of understanding environment and climate change effects.

The expose research is under peer review in a refereed journal:

(2010) **Supervised pre-processing approaches in multiple class-variables classification for fish recruitment forecasting**. *Fernandes J.A.*, Lozano J.A., Inza I., Irigoien X., Rodríguez J.D. and Pérez A. Applied Soft Computing. Submitted.

Part III

Conclusions and future work

7

Conclusions

The topic of this dissertation is the application of supervised classification methods in fisheries management related activities. Fisheries management is a multi-disciplinary area, where biological and the methodological issues from mathematics, statistics or computer science have to be appropriately considered. Although many of the methods presented in this dissertation are not new for the machine learning community, many are relatively unknown in marine science. Indeed, there is uncertainty on the methods to use as well as the correct way of using them in their practical application in real domains related to fisheries management. Therefore, many necessary issues have not been covered in the literature so far about the practical application of such methods. The main contributions of this dissertation are concerned about achieving a good trade off between appropriate dealing with methodological and biological issues.

The relevance of an appropriate selection of a pipeline of methods (not only the model induction algorithms), depending on the specific objectives of each study, is shown in different parts of this dissertation. In addition, other topics such as discretization, feature selection or pipeline evaluation are considered for both classical uni-dimensional classification approach and the novel multi-dimensional approach. In this dissertation we have applied the proposed methods mainly in zooplankton classification and fish recruitment forecasting.

The rest of this chapter is organised as follows. Section 7.1 summarises the contributions of this dissertation in the application of supervised classification methods to fisheries management, in particular in the new paradigm of multi-dimensional classification for simultaneous forecasting of multiple fish species. In Section 7.2 the list of publications obtained during the last four years of PhD work are provided. Finally, Section 7.3 explores lines of work that this dissertation has left open.

7.1 Main contributions of the thesis

The main contributions have been focused on the topics of plankton classification and fish recruitment forecast. In order to accomplish these contributions, a set of methodological issues has been addressed considering the biological objectives. These methodologies include algorithms that cover three main aspects: class variable definition; proposal of pipelines of supervised classification methods; and proposals of supervised pre-processing methods for multiple class variables classification.

7.1.1 Zooplankton classification

The main contribution in this area has been in zooplankton classification, where the author has faced the task of providing experts with a tool that allows them to evaluate the trade-off between the number of classes and the performance in a training set (Fernandes et al., 2009c). The end-user can accept or reject mergers of classes, depending on whether they are ecologically meaningful and the objectives of the research. This method permits to balance both objectives: a) maximization of the number of classes; and b) performance, guided by the end-user. A wrapper machine learning method is proposed in order to accomplish this trade-off.

7.1.2 Fish recruitment forecasting

Significant contributions of this dissertation are in the domain of fish recruitment forecasting. The contributions have been mostly in the application of a set of well-known and novel techniques of supervised classification to recruitment forecast for fisheries management purposes.

Firstly, in Fernandes et al. (2010c) a methodological pipeline of suited machine learning methods is proposed. The proposed pipeline has several desired properties for fisheries management: a) forecast with its estimated *uncertainty* associated; b) forecast and scenarios easy to interpret; c) recruitment and the boundaries of the factors that can be interpreted; d) a high degree of factors stability; e) error distribution balanced through all recruitment levels and; f) robust and honest error estimation. A wrapper method is proposed for discretizing a continuous target variable (recruitment) with the objective of balancing the error distribution between the different class values in fish recruitment forecasting.

Secondly, the proposal and application of specific pre-processing methods for the novel *multi-dimensional classification* approach has been achieved. A whole methodological pipeline is proposed in order to forecast multiple species simultaneously (Fernandes et al., 2010b). The results show how this approach allows to improve not only the forecasts of each species, but also the chance of being right in all the species simultaneously. In addition, experimentation with synthetic datasets and a meta-learning process are performed in order

to identify not only the methods with superior behaviour, but also the circumstances under which each multi-dimensional pre-processing method can have a superior behaviour. Finally, proposals and adaptations of performance measures for the multi-dimensional approach are presented for *accuracy* and *Brier score* measures.

7.2 Other related contributions

In this section, a brief description of several additional contributions not described in this thesis is provided. These contributions are related with the methodologies proposed and the work performed during this thesis. Most of these additional contributions are collaborations with other authors.

7.2.1 Samples classification

The method proposed in this thesis (Fernandes et al., 2009c) has allowed to process thousand of zooplankton images that has led to the next contribution presented. This additional contribution discarded the zooplankton as a limitation or explanation for the low anchovy recruitment last years in the Bay of Biscay (Irigoien et al., 2009).

Other contributions related to classification of samples in the Bay of Biscay during this thesis work have been mostly in collaborations with other authors: a) phytoplankton classification (Zarauz et al., 2009; Denis et al., 2009; Alcaraz et al., 2010); b) zooplankton distribution (Zarauz, 2007; Zarauz et al., 2008; Bachiller, 2008; Bachiller et al., 2010; Bachiller, 2010); and, c) otolith age classification (Ascoreca et al., 2008).

7.2.2 Fish recruitment forecasting

In addition to the main work of this thesis, during the year 2009, a forecast of anchovy based on environmental factors was performed (Fernandes et al., 2009b) based on the methodology proposed in Fernandes et al. (2010c). A medium recruitment due to improvement of environmental conditions during the year 2009 was forecasted. In addition, the results of another study based on distribution of juveniles confirmed this forecast, leading to a limited opening of the fisheries. The use of some of the methodologies for a further establishment of critical levels of recruitment for management purposes is under current discussion. Its potential use for several other species (Goikoetxea, 2010) and its regular use in advising to government agencies and groups of interest (Ibaibarriaga et al., 2010; Andonegi E., 2010) is also being evaluated.

The need to perform continuous forecast in order to incorporate them in other modelling approaches led to studying the use of *naive for regression Bayesian network* modelling (Andonegi et al., 2010a,b). However, the factors

and recruitment boundaries remain those ones selected in Fernandes et al. (2010c), since the application of such a methodology managing continuous variables might have too high a CPU-time cost. Similarly, the use of *flexible Bayesian network classifiers* with continuous factors has also been studied (Fernandes et al., 2009a, 2010a) in order to provide a more detailed distribution of recruitment probabilities. These works are in the line of other works where these kinds of approaches are being applied to real domain problems (Fernández et al., 2007; Aguilera et al., 2010; Fernández et al., 2010).

7.2.3 Other domains

Finally, some contributions in the area of teledetection for habitat classification have been carried out in support of other researchers (Grande et al., 2009; Grande, 2009). There are also preliminary contributions to future works in acoustics, which are just beginning.

7.3 List of main publications and contributions

7.3.1 First author in Refereed JCR-Journals publications

- (2010) **Supervised pre-processing approaches in multiple class-variables classification for fish recruitment forecasting**. *Fernandes J.A.*, Lozano J.A., Inza I., Irigoien X., Rodríguez J.D. and Pérez A. Applied Soft Computing. Submitted.
- (2010) **Robust machine-learning techniques for recruitment forecasting of North East Atlantic fish species**. *Fernandes J.A.*, Irigoien X., Lozano J.A., Inza I. and Pérez A. ICES Journal of Marine Science. Submitted.
- (2010) **Fish recruitment prediction, using robust supervised classification methods**. *Fernandes J.A.*, Irigoien X., Goikoetxea N., Lozano J.A., Inza I., Pérez A. and Bode A. Ecological Modelling, 221(2): 338-352.
- (2009) **Optimizing the number of classes in automated zooplankton classification**. *Fernandes J.A.*, Irigoien X., Boyra G., Lozano J.A. and Inza I. Journal of Plankton Research 31(1): 19-29.

7.3.2 Colaborations in JCR-Journals publications

- (2010) **The potential use of a Gadget model to forecast stock responses to climate change in combination with Bayesian Networks: the case of the Bay of Biscay anchovy**. Andonegi E., *Fernandes J.A.*, Quincoces I., Uriarte A., Pérez A., Howell D. and Stefansson G. ICES Journal of Marine Science. Submitted.

- (2009) **Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment**. Irigoien X., *Fernandes J.A.*, Grosjean P., Denis K., Albaina A. and Santos M. Journal of Plankton Research 31(1): 1-17. Featured article.
- (2009) **Changes in plankton size structure and composition, during the generation of a phytoplankton bloom, in the central Cantabrian sea**. Zarauz L., Irigoien I. and Fernandes J.A. Journal of Plankton Research. 31(2): 193-207.
- (2008) **Modelling the influence of abiotic and biotic factors on plankton distribution in the Bay of Biscay, during three consecutive years (2004-06)**. Zarauz L., Irigoien X. and Fernandes J.A. Journal of Plankton Research 30(8): 857-872.

7.4 Future work

In the previous sections, several ot the author contributions and other lines of work with colaborators have been presented. However, there are many lines of potential interesting work the author has identified.

Starting with zooplankton classification and samples classification, more work to help the end-user in designing the training set is needed, as well as to deal with the imbalanced dataset problem and its ecological implications. The use of cost-sensitive classification in samples classification and unidimensional as well as multi-dimensional fish recruitment classification would be a valuable tool if properly addressed. Finally, the proposed methodologies can be applied to other domains in marine science and other biological sciences with significant results; in particular, the use of multi-dimensional classifiers for ecosystem approaches and water quality challenges. The success of all future work depends on the communication between different sciences to accomplish a trade-off between methodological developments and their proper application. Other novel approaches of interested are related with the use of optimization methods similarly to the work presented in Ermon et al. (2010).

Part IV

Appendices

A

Appendix: Mathematical notation

Following is a list of the most frequent mathematical notations in the thesis.

- C: Class variable or discrete target variable.
- Y: Continuous target variable.
- m: Number of class variables or dimensions.
- j: Class index.
- r: Number of class values in a class variable.
- o: Number of values of the class cartesian product.
- g: Index for number of values of the class cartesian product.
- l: Index for number of values of one class variable.
- r_j: the number of class values of the j class variable.
- r_o: the number of class values of the class cartesian product.
- r_s: the sum of number of class values in all the class variables.
- n: Number of features. It is also used to denote the number of time a *cross-validation* is repeated.
- X: Feature.
- i: Features index.
- $\mathbf{x}$: Unlabeled instance.
- S: Full dataset.
- N: Number of instances or cases.
- N_c: Subset of instances with the same class value.
- p_{kj}: Predicted probability of a class value.
- k: Cases index. It is also used to denote the number of folds in *cross-validation*.
- M: A random classifiers based on KBN.
- G: The structure of MN.
- y_{kj}: 1 if it is the observed class value, 0 otherwise.
- f_{kj}: non-observed class values (failures) within a combination of class variables being evaluated.
- z: Number of features in a subset.
- $\mathbf{pa}_i$: Parents of a feature in a structure.

- e: Kernel coordinates.
- q: Number of kernels.
- h: Kernel smooth parameter.
- H: A $n \times n$ bandwidth or smoothing matrix (BM).
- K_H: Gaussian kernel.
- t_{cx}: the average class-feature correlation in CFS.
- t_{xx}: the average feature-feature correlation in CFS.
- $\mathbf{pd}_i$: Discrete parents of a feature X_i.
- $\mathbf{pc}_i$: Continuous parents of a feature X_i.

B

Intrinsic data characteristics measured by means of Information Theory

A set of statistics, based upon Information Theory Cover and Thomas (2006), have been adapted to capture the main intrinsic data characteristics of multi-dimensional datasets. These statistics have been normalised to permit comparison between different datasets, which have a different number of variables and values.

Class entropy:
The normalized class entropy (CE) for the vector of classes is:

$$CE(C_1, ..., C_m) = \frac{1}{m} \sum_{j=1}^{m} \frac{H(C_j)}{\log_2 r_j},$$

where m is the number of class variables and r_j is the number of values of the j^{th} class variable.

Classes interaction:
The classes interaction (CI) is:

$$CI(C_1, ..., C_m) = \frac{2}{m(m-1)} \sum_{1 \leq i < j \leq m} \frac{2I(C_i; C_j)}{H(C_i) + H(C_j)}$$

which is calculated using the symmetrical *uncertainty* measure Witten and Frank (2005) for each pair of class variables.

Feature redundancy:
The features redundancy (FRd), taking all feature pairs into account, is:

$$FRd(X_1, ..., X_n) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \frac{I(X_i; X_j)}{\min(H(X_i), H(X_j))},$$

which is normalized since $I(X;Y) \leq min(H(X), H(Y)) \leq H(X,Y)$ Yao and Regina (2003).

Feature relevance:
The features (univariate) relevance (FRv), with respect to each class, is:

$$FRv(X_1, ..., X_n, C_1, ..., C_m) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{I(X_i; C_j)}{H(C_j)}$$

Features-class interaction:
The features-class interaction (FI), based upon the 3-way interaction metric Jakulin (2005) which measures the *uncertainty* shared by three variables, is the average interaction between variables triples (two features and one class variable):

$$FI(X_1, ..., X_n, C_1, ..., C_m) = \frac{2}{m * n(n-1)} \sum_{k=1}^{m} \sum_{1 \leq i < j \leq n} \frac{I(X_i; X_j; C_k)}{I(C_k; (X_i, X_j))}$$

where $I(X_i; X_j; C_k) = I(X_i; X_j|C) - I(X_i; X_j)$
and $I(C_k; (X_i, X_j)) = I(C_k; X_i) + I(C_k; X_j) + I(X_i; X_j; C_k)$.

Classes-feature interaction:
The classes-feature interaction (FCI), similar to features-class interaction, considers triples of two class variables and one feature, and it is given by:

$$FCI(X_1, ..., X_n, C_1, ..., C_m) = \frac{2}{n * m(m-1)} \sum_{k=1}^{n} \sum_{1 \leq i \leq j \leq m} \frac{I(C_i; C_j; X_k)}{I(X_k; (C_i, C_j))}$$

References

Abramowitz, M., Stegun, I., 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover publications, Washington, DC, USA.

Aguilera, P., Fernández, A., Reche, F., Rumí, R., 2010. Hybrid Bayesian network classifiers: Application to species distribution models. Environ. Modell. Softw. 25 (12), 1630–1639.

Alcaraz, M., Saiz, E., Calbet, A., Trepat, I., Broglio, E., 2003. Estimating zooplankton biomass through image analysis. Mar. Biol. 143 (2), 307–315.

Alcaraz, M., Saiz, E., Lebourges-Dhaussy, A., Graña, R., Cotano, U., Fernandes, J. A., Isari, S., Zamora, S. Mouriño, B., Irigoien, X., 2010. Small-scale vertical distribution of zooplankton in the Catalan Sea: Relationships with physical characteristics. In: Aquatic Sciences: Global Changes from the Center to the Edge. Rapp. Comm. Int. Mer Medit. Vol. 39 - pp. 84. 39th CIESM Congress. Venice, Italy.

Alheit, J., Hagen, E., 1997. Long-term climate forcing of European herring and sardine populations. Fish. Oceanogr. 6 (2), 130–139.

Ali, S., Smith, K. A., 2006. On learning algorithm selection for classification. Appl. Soft Comput. 6 (2), 119–138.

Allain, G., Petitgas, P., Lazure, P., 2001. The influence of mesoscale ocean processes on anchovy (Engraulis encrasicolus) recruitment in the Bay of Biscay estimated with a three-dimensional hydrodynamic mode. Fish. Oceanogr. 10 (2), 151–163.

Allain, G., Petitgas, P., Lazure, P., 2007. The influence of environment and spawning distribution on the survival of anchovy (Engraulis encrasicolus) larvae in the Bay of Biscay (NE Atlantic) investigated by biophysical simulations. Fish. Oceanogr. 16 (6), 506–514.

Allison, P., 2001. Missing data. Sage Publications.

Alpaydin, E., 2004. Introduction to Machine Learning. MIT Press, Cambridge, MA, USA.

Álvarez, P., Fives, J., Motos, L., Santos, M., 2004. Distribution and abundance of European hake Merluccius merluccius (L.), eggs and larvae in the

North East Atlantic waters in 1995 and 1998 in relation to hydrographic conditions. J. Plankton Res. 26 (7), 811.

Amaldi, E., Kann, V., 1998. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theor. Compt. Sci. 209 (1-2), 237–260.

Anderson, C., Hsieh, C., Sandin, S., Hewitt, R., Hollowed, A., Beddington, J., May, R., Sugihara, G., 2008. Why fishing magnifies fluctuations in fish abundance. Nature 452 (7189), 835–839.

Andonegi, E., Fernandes, J., Uriarte, A., Pérez, A., Howell, D., Stefansson, G., 2010a. The potential use of a Gadget model to predict stock responses to climate change in combination with Bayesian Networks: the case of the Bay of Biscay anchovy. ICES. J. Mar. Sci. Submitted.

Andonegi, E., Fernandes, J., Uriarte, A., Pérez, A., Howell, D., Stefansson, G., 2010b. The potential use of a Gadget model to predict stock responses to climate change in combination with Bayesian Networks: the case of the Bay of Biscay anchovy. In: International Symposium Climate Change Effects on Fish and Fisheries: Forecasting Impacts, Assessing Ecosystem Responses, and Evaluating Management Strategies. Sendai, Japan.

Andonegi E., Quincoces I., M. H. F. J. A. U. A. S. S. C. S. V. F. H. M. L. S. . P. P., 2010. UNCOVER: Fish stock recovery strategies - Report from the Bay of Biscay. Tech. rep., UNCOVER committee. Submitted.

Ascoreca, A., Fernandes, J. A., Cotano, U., Uriarte, A., Irigoien, X., 2008. Relevance of otolith features for anchovy age classification. In: Eleventh International Symposium on Oceanography of the Bay of Biscay, San Sebastian, Guipuzkoa, Spain.

Ashjian, C., Davis, C., Gallager, S., Alatalo, P., 2001. Distribution of plankton, particles, and hydrographic features across Georges Bank described using the Video Plankton Recorder. Deep-Sea Res. II 48 (1-3), 245–282.

Augustin, N., Borchers, D., Clarke, E., Buckland, S., Walsh, M., 1998. Spatiotemporal modelling for the annual egg production method of stock assessment using generalized additive models. Can. J. Fish. Aquat. Sci. 55 (12), 2608–2621.

Bachiller, E., 2008. Feeding behaviour and diet of anchovy juveniles (Engraulis encrasilous L.) in the Bay of Biscay. In: European Master of Science in Marine Enviroment and Resources, Leioa, Vizcaya, Spain.

Bachiller, E., 2010. Trophic Ecology of small pelagic species in the Bay of Biscay. Ph.D. thesis, University of the Basque Country, Leioa, Vizcaya, Spain.

Bachiller, E., Fernandes, J. A., Irigoien, X., 2010. A comparison between digital camera and scanner as imaging devices for semi-automated zooplanktton classification using microscope classification as control. In: Aquatic Sciences: Global Changes from the Center to the Edge. International Joint Meeting with ASLO & NABS, Santa Fe, NM, USA.

Bailey, K., Ciannelli, L., Bond, N., Belgrano, A., Stenseth, N., 2005. Recruitment of walleye pollock in a physically and biologically complex ecosystem: A new perspective. Prog. Oceanogr. 67 (1-2), 24–42.

Bakun, A., 1996. Patterns in the Ocean: Ocean Processes and Marine Population Dynamics. University of California Sea Grant, San Diego, California, USA, in cooperation with Centro de Investigaciones Biológicas del Noroeste, La Paz, Baja California Sur, México.

Banse, K., 1995. Zooplankton: Pivotal role in the control of ocean production: I. Biomass and production. ICES. J. Mar. Sci. 52 (3-4), 265.

Barnes, N., 2010. Publish your computer code: it is good enough. Nature 467 (753).

Barnston, A., Livezey, R., 1987. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Month. Weather Rev. 115 (6), 1083–1126.

Bartolino, V., Colloca, F., Sartor, P., Ardizzone, G., 2008. Modelling recruitment dynamics of hake, Merluccius merluccius, in the central Mediterranean in relation to key environmental variables. Fish. Res. 92 (2-3), 277–288.

Batista, G., Monard, M., 2003. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17 (5), 519–533.

Baumgartner, T., Soutar, A., Ferreira-Bartrina, V., 1992. Reconstruction of the history of Pacific sardine and northern anchovy populations over the past two millennia from sediments of the Santa Barbara Basin, California. CalCOFI Rep 33, 24–40.

Beare, D., Gislason, A., Astthorsson, O., McKenzie, E., 2000. Assessing long-term changes in early summer zooplankton communities around Iceland. ICES. J. Mar. Sci. 57 (6), 1545.

Bell, J., Hopcroft, R., 2008. Assessment of ZooImage as a tool for the classification of zooplankton. J. Plankton Res. 30 (12), 1351.

Bellan, G., 1967. Pollution et peuplements benthiques sur substrat meuble dans la region de Marseille. Rev. Intern. Oceanogr. Med. 8, 51–95.

Bellier, E., Planque, B., Petitgas, P., 2007. Historical fluctuations in spawning location of anchovy (Engraulis encrasicolus) and sardine (Sardina pilchardus) in the Bay of Biscay during 1967-73 and 2000-2004. Fish. Oceanogr. 16, 1–15.

Ben-Bassat, M., 1982. Use of distance measures, information measures and error bounds in feature evaluation. Handbook of Statistics 2, 773–791.

Benfield, M., Grosjean, P., Culverhouse, P., Irigoien, X., Sieracki, M., Lopez-Urrutia, A., Dam, H., Hu, Q., Davis, C., Hansen, A., Pilskaln, C., Riseman, E., Schultz, H., Utgoff, P., Gorsky, G., 2007. RAPID: research on automated plankton identification. Oceanogr. 20 (2), 172–187.

Benner, T., 1999. Central England temperatures: long-term variability and teleconnections. Int. J. Climatol. 19 (4), 391–403.

Bernardo, J., Smith, A., 2001. Bayesian theory. Meas. Sci. and Technol. 12, 221.

Beverton, R. J. H., Holt, S. J., 1957. On the dynamics of exploited fish populations. Fishery Invest., Ser. II, Vol. XIX. Ministry of Agriculture, Fisheries and Food.

Bezdek, J., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Pub., Norwell, MA, USA.

Bielza, C., Li, G., Larrañaga, P., 2010. Multi-dimensional classification with bayesian networks. Technical Report. Department of Artificial Intelligence, Polytechnic University of Madrid. UPM-FI/DIA/2010-1. Madrid, Spain.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.

Blanco, R., Inza, I., Larrañaga, P., 2003. Learning Bayesian networks in the space of structures by estimation of distribution algorithms. Int. J. Intell. Syst. 18 (2), 205–220.

Blaschko, M., Holness, G., Mattar, M., Lisin, D., Utgoff, P., Hanson, A., Schultz, H., Riseman, E., 2005. Automatic in situ identification of plankton. In: Seventh IEEE Workshops on Application of Computer Vision, Washington, DC, USA. Vol. 1.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. ACM, p. 100.

Bode, A., Alvarez-Ossorio, M. T., Cabana, J. M., Porteiro, C., Ruiz-Villarreal, M., Santos, M. B., Bernal, M., Valdes, L., Varela, M., 2006. Recent changes in the pelagic ecosystem of the Iberian Atlantic in the context of multidecadal variability. In: Proceedings of the ICES CM Theme Session C:07.

Borja, Á., Fontán, A., Sáenz, J., Valencia, V., 2008. Climate, oceanography, and recruitment: the case of the Bay of Biscay anchovy (Engraulis encrasicolus). Fish. Oceanogr. 17 (6), 477–493.

Borja, Á., Franco, J., Pérez, V., 2000. A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Mar. Pollut. Bull. 40 (12), 1100–1114.

Borja, Á., Uriarte, A., Valencia, V., Motos, L., Uriarte, A., 1996. Relationships between anchovy (Engraulis encrasicolus L.) recruitment and the environment in the Bay of Biscay. Sci. Mar. 60, 179–192.

Borja, Á., et al., 1998. Relationships between anchovy (Engraulis encrasicolus) recruitment and environment in the Bay of Biscay (1967–1996). Fish. Oceanogr. 7 (3-4), 375–380.

Botsford, L. W., Castilla, J. C., Peterson, C. H., 1997. The management of fisheries and marine ecosystems. Science 277 (5325), 509.

Bouckaert, R., 1995. Bayesian belief networks: from construction to inference. Ph.D. thesis, University of Utrecht, Utrecht, Netherlands.

Bouckaert, R., 2003. Choosing between two learning algorithms based on calibrated tests. In: International Conference on Machine Learning. Vol. 20. pp. 51–58.

Bouckaert, R. R., Frank, E., 2004. Evaluating the replicability of significance tests for comparing learning algorithms. Lect. Notes Artif. Int., 3–12.

Boyra, G., I. X. A. A., Arregi, I., 2005. Plankton Visual Analyser. GLOBEC International Newsletter (11), 9–10.

Brazdil, P., Soares, C., Da Costa, J., 2003. Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. Mach. Learn. 50 (3), 251–277.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA, USA.

Brier, G. W., 1950. Verification of forecasts expressed in terms of probability. Month. Weather Rev. 78 (1), 1–3.

Brunel, T., Boucher, J., 2007. Long-term trends in fish recruitment in the north-east Atlantic related to climate change. Fish. Oceanogr. 16 (4), 336–349.

Buckheit, J., Donoho, D., 1995. Wavelab and reproducible research. Wavelets and statistics, 55.

Buntine, W., 1991. Theory refinement on Bayesian networks. In: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence. Vol. 91. pp. 52–60.

Calvo, B., 2008. Positive Unlabelled Learning with Applications in Computational Biology. Ph.D. thesis, University of the Basque Country, San Sebastian, Guipuzkoa, Spain.

Calvo, B., Larrañaga, P., Lozano, J., 2007. Learning Bayesian classifiers from positive and unlabeled examples. Pattern Recog. Lett. 28 (16), 2375–2384.

Castillo, E., Gutiérrez, J., Hadi, A., 1997. Expert Systems and Probabilistic Network Models. Springer Verlag, New York, NY, USA.

Cestnik, B., Kononenko, I., Bratko, I., 1987. A knowledge elicitation tool for sophisticated users. In: Proceedings of the Second European Working Session on Learning. pp. 31–45.

Chambers, J., 1983. Graphical Methods for Data Analysis.

Chapelle, O., Schölkopf, B., Zien, A., 2006. Semi-Supervised Learning. MIT Press, Cambridge, MA, USA.

Cheeseman, P., Stutz, J., 1996. Bayesian classification (AutoClass): Theory and results. Advances in knowledge discovery and data mining, 153–180.

Chen, D., Ware, D., 1999. A neural network model for forecasting fish stock recruitment. Can. J. Fish. Aquat. Sci. 56 (12), 2385–2396.

Chickering, D., 1996. Learning Bayesian networks is NP-complete. Learning from data: Artificial intelligence and statistics V 112, 121–130.

Chickering, D., Geiger, D., Heckerman, D., 1994. Learning Bayesian networks is NP-hard. Microsoft Research, 94–17.

Chickering, D., Geiger, D., Heckerman, D., 1995. Learning Bayesian networks: Search methods and experimental results. In: Proceedings of Fifth Conference on Artificial Intelligence and Statistics. pp. 112–128.

Chow, C., Liu, C., 1968. Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory 14 (3), 462–467.

Christensen, N., Bartuska, A., Brown, J., Carpenter, S., D'Antonio, C., Francis, R., Franklin, J., MacMahon, J., Noss, R., Parsons, D., et al., 1996. The

report of the Ecological Society of America committee on the scientific basis for ecosystem management. Ecol. Appl. 6 (3), 665–691.

Cleveland, W., 1984. Elements of Graphing Data.

Consortium, E., 2002. Elvira: An environment for probabilistic graphical models. In: Proceedings of the First European Workshop on Probabilistic Graphical Models. pp. 222–230.

Cooper, G., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. 9 (4), 309–347.

Correa, M., Bielza, C., Pamies-Teixeira, J., 2009. Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. Expert Syst. Appl. 36 (3), 7270–7279.

Cover, T. M., Thomas, J. A., 2006. Elements of Information Theory. Wiley, New York, NY, USA.

Crawley, M., 2002. Statistical Computing: an Introduction to Data Analysis Using S-Plus. John Wiley & Sons, New York, NY, USA Inc.

Culverhouse, P., Simpson, R., Ellis, R., Lindley, J., Williams, R., Parsini, T., Reguera, B., Bravo, I., Zoppoli, R., Earnshaw, G., et al., 1996. Automatic classification of field-collected dinoflagellates by artificial neural network. Mar. Ecol. Prog. Ser. 139 (1-3), 281–287.

Culverhouse, P., Williams, R., Benfield, M., Flood, P., Sell, A., Mazzocchi, M., Buttino, I., Sieracki, M., 2006. Automatic image analysis of plankton: future perspectives. Mar. Ecol. Prog. Ser. 312, 297–309.

Culverhouse, P., Williams, R., Reguera, B., Herry, V., González-Gil, S., 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Mar. Ecol. Prog. Ser. 247, 17–25.

Cushing, D., 1982. Climate and Fisheries. Academic Press, London, UK.

Davis, C., 2008. Optical imaging of ocean plankton: a fantastic voyage. In: Digital Holography and Three-Dimensional Imaging. Optical Society of America.

Davis, C., Gallager, S., Berman, M., Haury, L., Strickler, J., 1992. The Video Plankton Recorder(VPR): Design and Initial Results. Ergebnisse der Limnologie 6, 36.

Davis, C., Hu, Q., Gallager, S., Tang, X., Ashjian, C., 2004. Real-time observation of taxa-specific plankton distributions: an optical sampling method. Mar. Ecol. Prog. Ser. 284, 77–96.

Davis, C., Thwaites, F., Gallager, S., Hu, Q., 2005. A three-axis fast-tow digital Video Plankton Recorder for rapid surveys of plankton taxa and hydrography. Limnol. Oceanogr.: Methods 3, 59–74.

De Campos, L., Fernández-Luna, J., Gámez, J., Puerta, J., 2002. Ant colony optimization for learning Bayesian networks. International Journal of Approximate Reasoning 31 (3), 291–311.

De Campos, L., Puerta, J., 2001. Stochastic local and distributed search algorithms for learning belief networks. In: Proceedings of the Third International Symposium on Adaptive Systems: Evolutionary Computation and Probabilistic Graphical Model. pp. 109–115.

De Oliveira, J., Uriarte, A., Roel, B., 2005. Potential improvements in the management of Bay of Biscay anchovy by incorporating environmental indices as recruitment predictors. Fish. Res. 75 (1-3), 2–14.

de Waal, P. R., van der Gaag, L. C., 2007. Inference and learning in multi-dimensional Bayesian network classifiers. Lect. Notes Artif. Int. 4724, 501–511.

Delavallade, T., Dang, T. H., 2007. Using entropy to impute missing data in a classification task. In: Proceedings of the IEEE International Conference on Fuzzy Systems. Vol. 7.

Dempster, A., Laird, N., Rubin, D., et al., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statistical Society, Series B 39 (1), 1–38.

Denis, K., Tunin-Ley, A., Fernandes, J., Maurer, S., Parent, J., Belin, C., Irigoien, X., Grosjean, P., 2009. FlowCAM/PhytoImage intercalibration exercise. In: Third SCOR WG130 meeting, Baton Rouge, LA, USA.

Deḿsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.

Dietterich, T., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10 (7), 1895–1923.

Domingos, P., 1999. The role of Occam's razor in knowledge discovery. Data Min. Knowl. Disc 3 (4), 409–425.

Domingos, P., Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. 29 (2), 103–130.

Dominguez-Caballero, J., Loomis, N., Li, W., Hu, Q., Milgram, J., Barbastathis, G., Davis, C., 2007. Advances in plankton imaging using digital holography. In: Adaptive Optics: Analysis and Methods/Computational Optical Sensing and Imaging/Information Photonics/Signal Recovery and Synthesis Topical Meetings on CD-ROM.

Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: A. Prieditis and S. Russell, (eds.), Proceedings of International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, USA, pp. 194–202.

Dreyfus-León, M., Chen, D. G., 2007. Recruitment prediction with genetic algorithms with application to the Pacific Herring fishery. Ecol. Model. 203 (1-2), 141–146.

Dreyfus-León, M., Schweigert, J., 2008. Recruitment prediction for Pacific herring (Clupea pallasi) on the west coast of Vancouver Island, Canada. Ecol. Inform. 3 (2), 202–206.

Drummond, C., Holte, R., 2000. Exploiting the Cost (In) sensitivity of Decision Tree Splitting Criteria. In: Proceedings of the Seventeenth International Conference on Machine Learning. pp. 239–246.

Duarte, C., 2007. Marine ecology warms up to theory. Trends Ecol. Evol. 22 (7), 331–333.

Duda, R. O., Hart, P. E., 1973. Pattern Classification and Scene Analysis. John Willey & Sons, New York, NY, USA.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification. Wiley, New York, NY, USA.

Edwards, S. F., Link, J. S., Rountree, B. P., 2004. Portfolio management of wild fish stocks. Ecol. Econ. 49 (3), 317–329.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7 (1), 1–26.

Ermon, S., Conrad, J., Gomes, C., Selman, B., 2010. Playing games against nature: optimal policies for renewable resource allocation. In: Grünwald, P., Spirtes, P. (Eds.), Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. AUAI Press.

Essington, T. E., 2001. The precautionary approach in fisheries management: the devil is in the details. Trends Ecol. Evol. 16 (3), 121–122.

Etxeberria, R., Larrañaga, P., Picaza, J., 1997. Analysis of the behaviour of genetic algorithms when learning Bayesian network structure from data. Pattern Recog. Lett. 18 (11-13), 1269–1273.

Fayyad, U., Irani, K., 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence. pp. 1022–1027.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: An overview. In: Fayyad, U.M., et al. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI/MIT. American Association for Artificial Intelligence Menlo Park, CA, USA, pp. 1–24.

Fernandes, J., Irigoien, X., Goikoetxea, N., Uriarte, A., Lozano, J., Inza, I., Pérez, A., 2009a. Robust approaches to supervised machine learning techniques for seven fish species recruitment prediction in fisheries. In: ICES/PICES/UNCOVER Symposium on Rebuilding Depleted Fish Stocks - Biology, Ecology, Social Science and Management Strategies. Warnemünde/Rostock, Germany.

Fernandes, J., Irigoien, X., Lozano, J., Inza, I., Pérez, A., Goikoetxea, N., 2010a. Robust machine-learning techniques for recruitment forecasting of North East Atlantic fish species. ICES. J. Mar. Sci. Submitted.

Fernandes, J., Irigoien, X., Uriarte, A., Ibaibarriaga, L., Lozano, J., Inza, I., 2009b. Anchovy Recruitment Mixed Long Series prediction using supervised classification. Tech. rep., Working document to the ICES benchmark workshop on short lived species (WKSHORT), Bergen, Norway.

Fernandes, J., Lozano, J., Inza, I., Irigoien, X., Rodríguez, J., Pérez, A., 2010b. Supervised pre-processing approaches in multiple class-variables classification for fish recruitment forecasting. Appl. Soft Comput. Submitted.

Fernandes, J. A., Irigoien, X., Boyra, G., Lozano, J. A., Inza, I., 2009c. Optimizing the number of classes in automated zooplankton classification. J. Plankton Res. 31 (1), 19–29.

Fernandes, J. A., Irigoien, X., Goikoetxea, N., Lozano, J. A., Inza, I., Pérez, A., Bode, A., 2010c. Fish recruitment prediction, using robust supervised classification methods. Ecol. Model. 221 (2), 338–352.

Fernández, A., Morales, M., Salmerón, A., 2007. Tree augmented naive Bayes for regression using mixtures of truncated exponentials: application to higher education management. In: Proceedings of the 7th International Conference on Intelligent Data Analysis. pp. 59–69.

Fernández, A., Nielsen, J., Salmerón, A., 2010. Learning Bayesian networks for regression from incomplete databases. Int. J. Uncertain. Fuzz. (18), 69–86.

Fiksen, Ø., Utne, A., Aksnes, D., Eiane, K., Helvik, J., Sundby, S., 1998. Modelling the influence of light, turbulence and ontogeny on ingestion rates in larval cod and herring. Fish. Oceanogr. 7 (3-4), 355–363.

Fix, E., Hodges, J., 1951. Discriminatory Analysis-Nonparametric Discrimination: Consistency properties. Tech. rep., USAF School of Aviation and Medicine, Randolph Field, 4.

Fogel, D., 2006. Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. Wiley-IEEE Press,New York, NY, USA.

Forgy, E., 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. Biometrics 21 (3), 768.

Fortier, L., Villeneuve, A., 1996. Cannibalism and predation on fish larvae by larvae of Atlantic mackerel, Scomber scombrus: trophodynamics and potential impact on recruitment. Fish. Bull. 94 (2), 268–281.

Fox, J., 2002. An R and S-Plus Companion to Applied Regression. Sage Publications, Thousand Oaks, CA, USA.

Francis, R. I. C., 2006. Measuring the strength of environment-recruitment relationships: the importance of including predictor screening within cross-validations. ICES J. Mar. Sci. 63 (4), 594.

Frank, E., Trigg, L., Holmes, G., Witten, I. H., 2000. Naive Bayes for regression. Mach. Learn. 41 (1), 5–15.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Mach. Learn. 29 (2), 131–163.

Fulton, E. A., Smith, A. D. M., Johnson, C. R., 2003. Effect of complexity on marine ecosystem models. Mar. Ecol. Prog. Ser. 253, 1–16.

García, S., Herrera, F., 2008. An extension on 'statistical comparisons of classifiers over multiple data sets' for all pairwise comparisons. J. Mach. Learn. Res. 9, 2677–2694.

Gaston, K., O'Neill, M., 2004. Automated species identification: why not? Philos. Trans. R. Soc. Lond. B Biol. Sci. 359 (1444), 655.

Geiger, D., 1992. An entropy-based learning algorithm of Bayesian conditional trees. In: Proceedings of the eighth conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, USA, pp. 92–97.

Geisser, S., 1975. The predictive sample reuse method with applications. J. Am. Stat. Assoc., 320–328.

Geurts, P., Wehenkel, L., 2000. Investigation and reduction of discretization variance in decision tree induction. In: Proceedings of the Eleventh European Conference on Machine Learning. pp. 162–170.

Gislason, A., Silva, T., 2009. Comparison between automated analysis of zooplankton using ZooImage and traditional methodology. J. Plankton Res. 31 (12), 1505.

Glémarec, M., Hily, C., 1981. Perturbations apportées à la macrofaune benthique de la baie de Concarneau par les effluents urbains et portuaires. Acta Oceanol. Appl. 2, 139–150.

Goikoetxea, N., 2010. Influence of the northeastern Atlantic oceanometeorological variability on the northern hake (Merluccius merluccius), based on the last three-decadal period data (1978-2006). Ph.D. thesis, University of the Basque Country, Leioa, Vizcaya, Spain.

Graham, C., Ferrier, S., Huettman, F., Moritz, C., Peterson, A., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19 (9), 497–503.

Grall, J., Glémarec, M., 1997. Using biotic indices to estimate macrobenthic community perturbations in the Bay of Brest. Estuar. Coast. Shelf Sci. 44, 43–53.

Grande, M., 2009. Assessment of the discrimination potential of bathymetric LIDAR and multispectral imagery for intertidal and subtidal habitats. In: European Master of Science in Marine Enviroment and Resources, Leioa, Vizcaya, Spain.

Grande, M., Chust, G., Fernandes, J., Galparsoro, I., 2009. Assesment of the discrimination potential fo bathymetric LIDAR and multispectral imagery for intertidal and subtidal habitats. In: International Symposium on Remote Sensing of Environment, Stresa, Italy.

Gray, J., Waldichuk, M., Newton, A., Berry, R., Holden, A., Pearson, T., 1979. Pollution-Induced Changes in Populations [and Discussion]. Philos. Trans. R. Soc. Lond. B Biol. Sci. 286 (1015), 545–561.

Grosjean, P., Picheral, M., Warembourg, C., Gorsky, G., 2004. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. ICES J. Mar. Sci. 61 (4), 518.

Guisan, A., Edwards, T., et al., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157 (2-3), 89–100.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009.

Guisan, A., Zimmermann, N., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2-3), 147–186.

Guyon, I., Aliferis, C., Elisseeff, A., 2007. Causal feature selection. In: Computational Methods of Feature Selection. pp. 63–82.

Hagle, T. M., Glen, E. M., 1992. Goodness-of-fit measures for Probit and Logit. Am. J. Polit. Sci. 36 (3), 762–784.

Hall, M., Smith, L., 1997. Feature subset selection: a correlation based filter approach. In: Proceedings of the Fourth International Conference on Neural Information Processing and Intelligent Information Systems. pp. 855–858.

Hall, M. A., 1999. Correlation-based Feature Selection for Machine Learning. Ph.D. Thesis, Waikato University, Hamilton, New Zealand.

Hall, M. A., 2000. Correlation-based feature selection of discrete and numeric class machine learning. In: Proceedings of the Seventeenth International Conference on Machine Learning. pp. 359–366.

Harrison, P., Parsons, T., 2000. Fisheries Oceanography: an Integrative Approach to Fisheries Ecology and management. Fish and Aquatic Resources Series 4, Blackwell Science, Malden, MA, USA.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London, UK.

Hays, G., Richardson, A., Robinson, C., 2005. Climate change and marine plankton. Trends Ecol. Evol. 20 (6), 337–344.

Heckerman, D., 1995. A tutorial on learning with Bayesian networks. Tech. rep., Microsoft Research, mSR-TR-95-06.

Heckerman, D., Geiger, D., Chickering, D., 1995. Learning Bayesian networks: The combination of knowledge and statistical data. Mach. Learn. 20 (3), 197–243.

Herskovits, E., Cooper, G., 1990. Kutato: An entropy-driven system for construction of probabilistic expert systems from databases. In: Proceedings of the sixth conference on Uncertainty in Artificial Intelligence. pp. 54–62.

Hily, C., 1984. Variabilité de la macrofaune benthique dans les milieux hypertrophiques de la Rade de Brest. Ph.D. thesis, University of Bretagne Occidentale, Brest, France.

Holland, J., 1992. Adaptation in Natural and Artificial Systems. MIT Press, Cambridge, MA, USA.

Hollowed, A. B., Bax, N., Beamish, R., Collie, J., Fogarty, M., Livingston, P., Pope, J., Rice, J. C., 2000. Are multispecies models an improvement on single-species models for measuring fishing impacts on marine ecosystems? ICES J. Mar. Sci. 57 (3), 707–719.

Hosmer, D., Lemeshow, S., 1989. Applied logistic regression. John Wiley & Sons, New York, NY, USA.

Hu, Q., Davis, C., 2006. Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. Mar. Ecol. Prog. Ser. 306, 51–61.

Hua, J., Temble, W. D., Dougherty, E. R., 2009. Performance of feature-selection methods in the classification of high-dimensional data. Pattern Recogn. 42, 409–424.

Hurrell, J., Kushnir, Y., Ottersen, G., Visbeck, M., 2003. An overview of the North Atlantic oscillation. Geophysical monograph, American Geophysical Union 134, 1–36.

Ibaibarriaga, L., Fernandez, C., Uriarte, A., Roel, B., 2008. A two-stage biomass dynamic model for Bay of Biscay anchovy: a Bayesian approach. ICES J. Mar. Sci. 65 (2), 191.

Ibaibarriaga, L., Irigoien, X., Santos, M., Motos, L., Fives, J., Franco, C., de LanzÓs, A., Acevedo, S., Bernal, M., Bez, N., et al., 2007. Egg and

larval distributions of seven fish species in north-east Atlantic waters. Fish. Oceanogr. 16 (3), 284–293.

Ibaibarriaga, L., Uriarte, A., Sanchez, S., Fernandes, J. A., Irigoien, X., 2010. Use of juvenile abundance indices for the management of the Bay of Biscay. Tech. rep., Working document to WGANSA, Lisbon, Portugal.

ICES, 2007. Report of the ICES/GLOBEC Workshop on Long-term Variability in SW Europe. ICES CM 02.

Inza, I., Larrañaga, P., Etxeberria, R., Sierra, B., 2000. Feature subset selection by Bayesian network-based optimization. Artif. Intell. 123 (1-2), 157–184.

Inza, I., Larrañaga, P., Sierra, B., Etxeberria, R., Lozano, J. A., Peña, J. M., 1999. Representing the behaviour of supervised classification learning algorithms by Bayesian networks. Pattern Recog. Lett. 20 (11-13), 1201–1209.

Irigoien, X., 2006. Reply to Horizons Article'Castles built on sand: dysfunctionality in plankton models and the inadequacy of dialogue between biologists and modellers' Flynn (2005). Shiny mathematical castles built on grey biological sands. J. Plankton Res. 28 (10), 965.

Irigoien, X., Fernandes, J., Grosjean, P., Denis, K., Albaina, A., Santos, M., 2009. Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment. J. Plankton Res. 31 (1), 1–17.

Irigoien, X., Fiksen, Ø., Cotano, U., Uriarte, A., Alvarez, P., Arrizabalaga, H., Boyra, G., Santos, M., Sagarminaga, Y., Otheguy, P., et al., 2007. Could Biscay Bay Anchovy recruit through a spatial loophole? Prog. Oceanogr. 74 (2-3), 132–148.

Irigoien, X., Grosjean, P., Urrutia, L., 2005. Report of the GLOBEC / SPACC workshop on Image Analysis to Count and Identify Zooplankton, San Sebastian, Guipuzkoa, Spain.

Irigoien, X., Harris, R., Verheye, H., Joly, P., Runge, J., Starr, M., Pond, D., Campbell, R., Shreeve, R., Ward, P., et al., 2002. Copepod hatching success in marine ecosystems with high diatom concentrations. Nature 419 (6905), 387–389.

Irigoien, X., Huisman, J., Harris, R., 2004. Global biodiversity patterns of marine phytoplankton and zooplankton. Nature 429 (6994), 863–867.

Isaacson, D., Madsen, R., 1985. Markov Chains-Theory and Applications. Tobert E. Kreiger Pub., Malabar, FL, USA.

Jain, A., Murty, M., Flynn, P., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323.

Jakulin, A., 2005. Machine Learning Based on Attribute Interactions. PhD. Thesis, University of Ljubljana, Slovenia.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study. Intell. Data Anal. 6 (5), 429–449.

Jardine, N., Sibson, R., 1971. Mathematical taxonomy. New York, NY, USA.

Jensen, F., Jensen, F., 1996. An Introduction to Bayesian Networks. UCL press, London, UK.

Jensen, F., Nielsen, T., 2001. Bayesian Networks and Decision Graphs. Springer-Verlag, New York, NY, USA.

Jin, J., 2009. Impossibility of successful classification when useful features are rare and weak. Proceedings of the National Academy of Sciences 106 (22), 8859.

John, G. H., Langley, P., 1995. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Vol. 1. pp. 338–345.

Jordan, M., 1998. Learning in Graphical Models. Kluwer Academic Pub., Norwell, MA, USA.

Kalousis, A., Gama, J., Hilario, M., 2004. On data and algorithms: Understanding inductive performance. Mach. Learn. 54 (3), 275–312.

Kalousis, A., Prados, J., Hilario, M., 2005. Stability of Feature Selection Algorithms. In: Fifth IEEE International Conference on Data Mining. pp. 218–225.

Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. 12 (1), 95–116.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence,. Vol. 14. pp. 1137–1145.

Kohavi, R., John, G. H., 1997. Wrappers for feature subset selection. Artif. Intell. 97 (1-2), 273–324.

Kohavi, R., Sommerfield, D., 1995. Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining. pp. 192–197.

Kononenko, I., 1995. On biases in estimating multi-valued attributes. In: Proceedings of the First International Joint Conference on Artificial Intelligence. Vol. 14. pp. 1034–1040.

Kononenko, I., Bratko, I., Roskar, E., 1984. Experiments in automatic learning of medical diagnostic rules. In: International School for the Synthesis of Experts Knowledge Workshop, Bled, Slovenia.

Korb, K., Nicholson, A., 2004. Bayesian Artificial Intelligence. CRC Press, Boca Raton, FL, USA.

Kotsiantis, S., 2007. Supervised Machine Learning: A Review of Classification Techniques. Inform. 31, 249–268.

Kotsiantis, S. B., Kanellopoulos, D., Pintelas, P. E., 2006. Data preprocessing for supervised leaning. Int. J. Comput. Sci. 1 (2), 1306–4428.

Kuncheva, L., 2007. A stability index for feature selection. In: Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications table of contents. pp. 390–395.

Lachenbruch, P., Mickey, M., 1968. Estimation of error rates in discriminant analysis. Technometrics, 1–11.

Langley, P., Iba, W., Thompson., K., 1992. An analysis of Bayesian classifiers. In: Proceedings of the 10th National Conference on Artificial Intelligence. pp. 223–228.

Laplace, P., 1912. Théorie analytique des probabilités.Paris: Courcier. Reprinted as Oeuvres Complétes de Laplace 7, 18781912. Paris: Gauthier-Villars.

Larrañaga, P., Kuijpers, C., Murga, R., Yurramendi, Y., 1996. Learning Bayesian network structures by searching for the best ordering with genetic algorithms. IEEE Trans. Syst., Man, Cybern. A, Syst., Humans 26 (4), 487–493.

Larrañaga, P., Lozano, J. A., Peña, J. M., Inza, I., 2005. Special issue on probabilistic graphical models for classification. Guest Editors: Mach. Learn. 59 (3), 211–212.

Larson, S. C., 1931. The shrinkage of the coefficient of multiple correlation. J. Educ. Psychol. 22 (1), 45–55.

Lauritzen, S., 1996. Graphical Models. Oxford University Press, New York, NY, USA.

Le Quéré, C., Harrison, S., Prentice, I., Buitenhuis, E., Aumont, O., Bopp, L., Claustre, H., Da Cunha, L., Geider, R., Giraud, X., et al., 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Glob. Change Biol. 11 (11), 2016–2040.

Lean, J., Beer, J., Bradley, R., 1995. Reconstruction of solar irradiance since 1610: Implications for climate change. Geophys. Res. Lett. 22 (23), 3195–3198.

Legendre, L., Le Fevre, J., Laval, U., de Bretagne Occidentale, U., 1991. From individual plankton cells to pelagic marine ecosystems and to global biogeochemical cycles. Particle analysis in oceanography 7, 261–300.

Leggett, W. C., Deblois, E., 1994. Recruitment in marine fishes: is it regulated by starvation and predation in the egg and larval stages? Neth. J. Sea Res. 32 (2), 119–134.

Little, R., Rubin, D., 2002. Statistical Analysis with Missing Data. Wiley Ser. Probab. Stat. Wiley, Hoboken, NJ, USA.

Liu, H., Motoda, H., 1998. Feature Extraction, Construction and sSelection: A Data Mining Perspective. Kluwer Academic Pub., Norwell, MA, USA.

Longhurst, A., 1991. Role of the marine biosphere in the global carbon cycle. Limnol. Oceanogr. 36 (8), 1507–1526.

Loomis, N., Dominguez-Caballero, J., Li, W., Hu, C., Davis, C., Milgram, J., Barbastathis, G., 2007. A compact, low-power digital holographic imaging system for automated plankton taxonomical classification. In: Fourth International Zooplankton Production Symposium–Human and Climate Forcing of Zooplankton Populations. Vol. 28.

Lucas, P., 2004. Restricted Bayesian network structure learning. Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing 49 (1), 217–232.

Luo, T., Kramer, K., Goldgof, D., Hall, L., Samson, S., Remsen, A., Hopkins, T., 2005. Active Learning to Recognize Multiple Types of Plankton. J. Mach. Learn. Res. 6, 589–613.

MacArthur, R., Wilson, E., 1967. The theory of island biogeography, 203 pp. Princeton University Press, Princeton, NJ, USA.

Mackas, D., 1984. Spatial autocorrelation of plankton community composition in a continental shelf ecosystem. Limnol. Oceanogr. 29 (3), 451–471.

MacKay, D., 2003. Information Theory, Iinference, and Learning Algorithms. Cambridge university press, New York, NY, USA.

MacKenzie, B., Horbowy, J., Koster, F., 2008. Incorporating environmental variability in stock assessment: predicting recruitment, spawner biomass, and landings of sprat (Sprattus sprattus) in the Baltic Sea. Can. J. Fish. Aquat. Sci. 65 (7), 1334–1341.

Mantyniemi, S., Kuikka, S., Rahikainen, M., Kell, L., Kaitala, V., 2009. The value of information in fisheries management: North Sea herring as an example. ICES. J. Mar. Sci. 66 (10), 2278.

Mattar, M., Murtagh, S., Hanson, A., 2009. Software Tools for Image Analysis. Tech. rep., Technical Report UM-CS-2009-017, Dept. of Computer Science, University of Massachusetts, Amherst, MA, USA.

Maury, O., 2010. An overview of APECOSM, a spatialized mass balanced. Prog. Oceanogr. 84 (1-2), 113–117.

McAllister, M. K., Ianelli, J. N., 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54 (2), 284–300.

McCulloch, W., Pitts, W., 1943a. A logical calculus of the ideas imminent in nervous activity. Bull. Math. Biophys. 5 (6), 115–133.

McCulloch, W. S., Pitts, W., 1943b. A logical calculus of the ideas immanent in nervous activity. B. Math. Biophys. 5, 115–137.

McFarlane, G., King, J., Beamish, R., 2000. Have there been recent changes in climate? Ask the fish. Prog. Oceanogr. 47 (2-4), 147–169.

McLeod, K., Leslie, H., 2009. Ecosystem-Based Management for the Oceans. Island Press, Washington, DC, USA.

Meiners, C. G., 2007. Importancia de la variabilidad climática en las pesquerías y biología de la merluza europea Merluccius merluccius (Linnaeus, 1758) de la costa Noroccidental Africana. Ph.D. thesis, Universidad Politécnica de Cataluña, Barcelona, Cataluña, Spain.

Meyer, R., Millar, R. B., 1999. Bayesian stock assessment using a state-space implementation of the delay difference model. Can. J. Fish. Aquat. Sci. 56 (1), 37–52.

Minsky, M., 1961. Steps toward artificial intelligence. Proceedings of the Institute of Radio Engineers 49 (1), 8–30.

Montgomery, D., Peck, E., 1992. Introduction to Linear Regression Analysis. John Wiley, New York, NY, USA.

Mosteller, F., Tukey, J. F., 1968. Data Analysis, Including Statistics. In G. Lindzey and E. Aronson, editors. Handbook of Social Psychology, Vol. II. Addison-Wesley, Reading, MA, USA.

Motos, L., 1996. Reproductive biology and fecundity of the Bay of Biscay anchovy population (Engraulis encrasicolus L.). Scientia Marina 60, 195–207.

Motos, L., Uriarte, A., Valencia, V., 1996. The spawning environment of the Bay of Biscay anchovy (Engraulis encrasicolus L.). Scientia Marina 60, 117–140.

Motos, L., Wilson, D., 2006. The Knowledge Base for Fisheries Management. Elsevier Science, Amsterdam, Holland.

Myers, J., Laskey, K., Levitt, T., 1999. Learning Bayesian networks from incomplete data with stochastic search algorithms. In: Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence. pp. 476–485.

Myers, R., Bridson, J., Barrowman, N., 1995. Summary of worldwide spawner and recruitment data. Can. Tech. Rep. Fish. Aquat. Sci., Canada.

Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Mach. Learn. 52 (3), 239–281.

Neapolitan, R., 2003. Learning Bayesian Networks. Pearson Prentice Hall, Upper Saddle River, NJ, USA.

Newman, K. B., Buckland, S. T., Lindley, S. T., Thomas, L., Fernández, C., 2006. Hidden process models for animal population dynamics. Ecol. Appl. 16 (1), 74–86.

Nilsson, N., 1965. Learning Machines. McGraw-Hill, New York, NY, USA.

O'Brien, R., 2004. Spatial decision support for selecting tropical crops and forages in uncertain environments. PhD thesis. Curtin University of Technology, Perth, Australia.

O'Brien, R., Cook, S., Peters, M., Corner, R., Mulla, D., 2004. A Bayesian Modeling Approach to Site Suitability Under Conditions of Uncertainty. In: Seventh International Conference on Precision Agriculture and Other Precision Resources Management, Minneapolis, MN, USA.

Olson, R., Sosik, H., 2007. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. Limnol. Oceanogr.: Methods 5, 195–203.

Pérez, A., Larrañaga, P., Inza, I., 2006. Information theory and classification error in probabilistic classifiers. In: Proceedings of the Ninth International Conference on Discovery Science. Lecture Notes in Artificial Intelligence. Vol. 4265. pp. 347–351.

Pazzani, M., 1996. Searching for dependencies in Bayesian classifiers. Learning from data: Artificial intelligence and statistics V, 239–248.

Pearl, J., 1985. Bayesian networks: A model of self-activated memory for evidential reasoning. In: Proceedings of the 7th Conference of the Cognitive Science Society. pp. 329–334.

Pearl, J., 1988. Probabilistic Reasoning in Intelligence Systems. Springer series in Statistics. Morgan Kaufman, San Francisco, CA, USA.

Pearson, T., Rosenberg, R., 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanogr. Mar. Biol. Ann. Rev 16, 229–311.

Pellet, J. P., Elisseeff, A., 2008. Using markov blankets for causal structure learning. J. Mach. Learn. Res. 9, 1295–1342.

Peña, J. M., Nilsson, R., Björkegren, J., Tegnér, J., 2007. Towards scalable and data efficient learning of Markov boundaries. Int. J. Approx. Reason. 45 (2), 211–232.

Pérez, A., 2010. Supervised classification in continuous domains with Bayesian networks. Ph.D. thesis, University of the Basque Country, San Sebastian, Guipuzkoa, Spain.

Pérez, A., Larrañaga, P., Inza, I., 2006. Information theory and classification error in probabilistic classifiers. Lect. Notes Comput. Sc. 4265, 347–351.

Pérez, A., Larrañaga, P., Inza, I., 2009. Bayesian classifiers based on kernel density estimation: Flexible classifiers. Int. J. Approx. Reason. 50 (2), 341–362.

Pianka, E., 1970. On r-and K-selection. Am. Nat. 104, 592.

Planque, B., Bellier, E., Lazure, P., 2007. Modelling potential spawning habitat of sardine (Sardina pilchardus) and anchovy (Engraulis encrasicolus) in the Bay of Biscay. Fish. Oceanogr. 16 (1), 16–30.

Planque, B., Buffaz, L., 2008. Quantile regression models for fish recruitment-environment relationships: four case studies. Mar. Ecol. Progr. Ser. 357, 213–223.

Provost, F., 2000. Machine learning from imbalanced data sets 101. In: Proceedings of the AAAI2000 Workshop on Imbalanced Data Sets.

Quinlan, J., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA, USA.

Quinn, G., Keough, M., 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge, UK.

Ragozin, D. L., Brown Jr, G., 1985. Harvest policies and nonmarket valuation in a predator-prey system. J. Environ. Econ. Manage. 12 (2), 155–168.

Rayner, N., Brohan, P., Parker, D., Folland, C., Kennedy, J., Vanicek, M., Ansell, T., Tett, S., 2006. Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 dataset. J. Clim. 19, 446–469.

Reid, G., 1987. Influence of solar variability on global sea surface temperatures. Nature 329 (6135), 142–143.

Reid, G., 1991. Solar total irradiance variations and the global sea surface temperature record. Journal of Geophysical Research 96, 2835–2844.

Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. J. Mach. Learn. Res. 3, 1371–1382.

Revoredo, K., Zaverucha, G., 2004. Search-Based Class Discretization for Hidden Markov Model for Regression. Lect. Notes Comput. Sci., 317–325.

Ricker, W. E., 1954. Stock and recruitment. J. Fish. Res. Board Can. 11, 559–623.

Rissanen, J., 1978. Modeling by the shortest data description. Automatica 14, 465–471.

Rodríguez, J. D., Lozano, J. A., 2008. Multi-objective learning of multi-dimensional Bayesian classifiers. In: Proceedings of the Eighth International Conference on Hybrid Intelligent Systems. pp. 501–506.

Rodríguez, J. D., Lozano, J. A., 2010. Learning bayesian networks classifiers for multi-dimensional supervised classification problems by means of a multi-objective approach. Technical report, EHU-KZAA-TR-3-2010, University of Basque Country, San Sebastian, Guipuzkoa, Spain.

Rodríguez, J. D., Pérez, A., Lozano, J. A., 2010. Sensitivity analysis of k-fold cross-validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32 (3), 569–575.

Ross, S., 2007. Introduction to Probability Models. Academic Press, San Diego, CA, USA.

Ruiz, J., GonzÁlez-QuirÓs, R., Prieto, L., Navarro, G., 2009. A Bayesian model for anchovy (Engraulis encrasicolus): the combined forcing of man and environment. Fish. Oceanogr. 18 (1), 62–76.

Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517.

Sahami, M., 1996. Learning limited dependence Bayesian classifiers. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 335–338.

Samson, S., Hopkins, T., Remsen, A., Langebrake, L., Sutton, T., Patten, J., 2001. A system for high-resolution zooplankton imaging. IEEE J. Ocean. Eng. 26 (4), 671–676.

Schirripa, M. J., Colbert, J. J., 2006. Interannual changes in sablefish (Anoplopoma fimbria) recruitment in relation to oceanographic conditions within the California Current System. Fish. Oceanogr. 15 (1), 25–36.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.

Sebastiani, P., Abad, M. M., Ramoni, M. F., 2005. Bayesian Networks. In: Maimom O., Rokach O. (ed.). The data mining and knowledge discovery handbook. Springer, New York, NY, USA, pp.193-230.

See, J., Campbell, L., Richardson, T., Pinckney, J., Shen, R., Guinasso Jr, N., 2005. Combining new technologies for determination of phytoplankton community structure in the northern gulf of mexico. J. Phycol. 41 (2), 305–310.

Shannon, C., Weaver, W., 1963. The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL, USA.

Sheehan, D., Hrapchak, B., 1980. Theory and practice of histotechnology.

Shumway, S., 1990. A review of the effects of algal blooms on shellfish and aquaculture. J. World Aquacult. Soc. 21 (2), 65–104.

Sieracki, C., Sieracki, M., Yentsch, C., 1998. An imaging-in-flow system for automated analysis of marine microplankton. Mar. Ecol. Prog. Ser. 168 (1), 285–296.

Sieracki, M., Benfield, M., Hanson, A., Davis, C., Pilskaln, C., Checkley, D., Sosik, H., Ashjian, C., Culverhouse, P., Cowen, R., et al., 2010. Optical Plankton Imaging and Analysis Systems for Ocean Observation. Proceedings of OceanObs 09: Sustained Ocean Observations and Information for Society 2, 21–25.

Silverman, B. W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London, UK.

Smith, T., Reynolds, R., Livezey, R., Stokes, D., 1996. Reconstruction of historical sea surface temperatures using empirical orthogonal functions. J. Climate 9 (6), 1403–1420.

Solow, A., Davis, C., Hu, Q., 2001. Estimating the taxonomic composition of a sample when individuals are classified with error. Mar. Ecol. Prog. Ser. 216, 309–311.

Spiegelhalter, D., Lauritzen, S., 1990. Sequential updating of conditional probabilities on directed graphical structures. Networks 20 (5), 579–605.

Spirtes, P., Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. Soc. Sci. Comput. Rev. 9 (1), 62.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., Levy, S., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21 (5), 631–643.

Steele, J., 1989. The ocean landscape. Landscape Ecol. 3 (3), 185–192.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. Roy. Statistical Society, Series B 36.

Stratoudakis, Y., Bernal, M., Borchers, D., Borges, M., 2003. Changes in the distribution of sardine eggs and larvae off Portugal, 1985–2000. Fish. Oceanogr. 12 (1), 49–60.

Tague, N., 2005. The Quality Toolbox. American Society for Qualit, Milwaukee, WI, USA.

Taylor, H., Karlin, S., 1998. An Introduction to Stochastic Modeling. Academic Press, San Diego, CA, USA.

Thompson, R., 1992. Graphical models in applied multivariate statistics. J. Classif. 9 (1), 159–160.

Torgo, L., Gama, J., 1997. Search-based class discretization. In: Proceedings of the Ninth European Conference on Machine Learning. pp. 266–273.

Uriarte, A., Roel, B. A., Borja, A., Allain, G., O'Brien, C., 2002. Role of Environmental indices in determining the recruitment of the Bay of Biscay anchovy. ICES CM 25.

Uriarte, A., Roel, B. A., Borja, A., Allain, G., O'Brien, C., 2008a. Report of the Working Group on the Assessment of Southern Shelf Stocks of Hake. ICES CM 07.

Uriarte, A., Roel, B. A., Borja, A., Allain, G., O'Brien, C. M., 2008b. Report of the Working Group on the Anchovy. ICES CM 04.

Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Model. 203 (3-4), 312–318.

van der Gaag, L., Renooij, S., 2001. Evaluation scores for probabilistic networks. In: Proceedings of the thirteenth Belgium-Netherlands Conference on Artificial Intelligence. pp. 109–116.

van der Gaag, L. C., de Waal, P. R., 2006. Multi-dimensional Bayesian network classifiers. In: Proceedings of the Third European Workshop in Probabilistic Graphical Models. pp. 107–114.

van der Gaag, L. C., Renooij, S., Witteman, C. L. M., Aleman, B. M. P., Taal, B. G., 2002. Probabilities for a probabilistic network: a case study in oesophageal cancer. Artif. Intell. Med. 25 (2), 123–148.

Vapnik, V., 2000. The Nature of Statistical Learning Theory. Springer Verlag, New York, NY, USA.

Villasante, S., Do Carme García-Negro, M., González-Laxe, F., Rodríguez, G., 2010. Overfishing and the Common Fisheries Policy: (un)successful results from TAC regulation? Fish Fish., doi: 10.1111/j.1467-2979.2010.00373.x.

Wand, M. P., Jones, M. C., 1995. Kernel Smoothing. Monographs on Statistics and Applied Probability, Chapman & Hall, London, UK.

Whittaker, J., 2009. Graphical models in applied multivariate statistics.

Wiebe, P., Benfield, M., 2003. From the Hensen net toward four-dimensional biological oceanography. Prog. Oceanogr. 56 (1), 7–136.

Wilk, M., Gnanadesikan, R., 1968. Probability plotting methods for the analysis for the analysis of data. Biometrika 55 (1), 1.

Witten, I. H., Frank, E., 2005. Data mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, CA, USA.

Wood, S., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. J. R.I Statist. Soc. B 70 (3), 495–518.

Worm, B., Hilborn, R., Baum, J., Branch, T., Collie, J., Costello, C., Fogarty, M., Fulton, E., Hutchings, J., Jennings, S., et al., 2009. Rebuilding global fisheries. science 325 (5940), 578.

Yang, Y., Webb, G., 2009. Discretization for naive-Bayes learning: Managing discretization bias and variance. Mach. Learn. 74 (1), 39–74.

Yao, Y. Y., Regina, S., 2003. Information-theoretic measures for knowledge discovery and data mining. In: Entropy easures, Maximum Entropy Principle and Emerging Applications. Karmeshu (Ed.), Springer, Berlin, pp. 115–136.

Yee, T., Mitchell, N., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2 (5), 587–602.

Yeung, K. Y., Bumgarner, R. E., Raftery, A. E., 2005. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 21 (10), 2394–2402.

Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5, 1205–1224.

Zadeh, L., 1965. Fuzzy sets. Inform. Contrlo 8 (3), 338–353.

Zarauz, L., 2007. Description and modelling of plankton biomasa distribution in the Bay of Biscay by means of image analisis-based methods. Ph.D. thesis, University of the Basque Country, Leioa, Vizcaya, Spain.

Zarauz, L., Irigoien, X., Fernandes, J., 2008. Modelling the influence of abiotic and biotic factors on plankton distribution in the Bay of Biscay, during three consecutive years (2004-06). J. Plankton Res. 30 (8), 857.

Zarauz, L., Irigoien, X., Fernandes, J., 2009. Changes in plankton size structure and composition, during the generation of a phytoplankton bloom, in the central Cantabrian sea. J. Plankton Res. 31 (2), 193–207.

Zarauz, L., Irigoien, X., Urtizberea, A., Gonzalez, M., 2007. Mapping plankton distribution in the Bay of Biscay during three consecutive spring surveys. Mar. Ecol. Prog. Ser. 345, 27–39.

Zhang, H., 2004. The optimality of naive Bayes. In: V. Barr and Z. Markov, (eds.), Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference. pp. 562–567.

Zhang, S., Zhang, C., Yang, Q., 2003. Data preparation for data mining. Appl. Artif. Intell. 17 (5), 375–381.

Zhou, Z., 2003. Three perspectives of data mining. Artif. Intell. 143 (1), 139–146.

Zhu, X., 2006. Semi-supervised learning literature survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin, Madison, MN, USA.

Zuur, A., Ieno, E., Smith, G., 2007. Analysing Ecological Data. Springer Verlag and Business Media LLC, New York, NY, USA.