EURO-BASIN Training Workshop on Introduction to statistical modelling tools, for habitat models development: Model Validation Performance measures Models comparison WEKA: open source software for data mining

Jose A. Fernandes

University of East Anglia / CEFAS
AZTI-Tecnalia
Intelligent Systems Group (University of Basque Country)

Outline

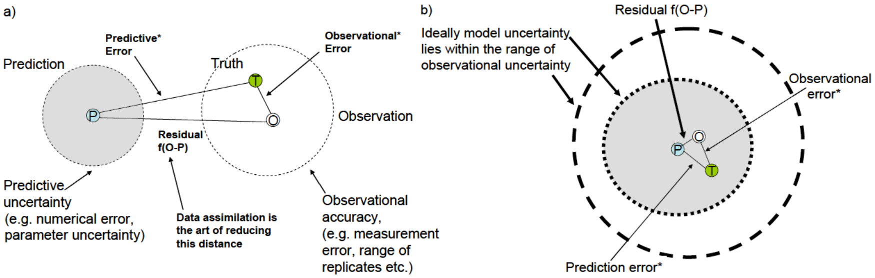
Outline

1. Model validation
2. Performance measures or metrics
   - Metrics in numeric prediction
   - Metrics in classification
3. Comparing methodologies and models
4. Examples
5. Weka: open source data mining tool
6. References

Introduction

- Slides based mainly in Witten and Frank (2005); Pérez et al. (2005); Allen (2009); Fernandes (2011)
- Objective: to measure how well a model represents truth.
- Truth cannot be accurately measured: observations.
- Questions:
  - How well the model fits the observations (goodness-of-fit)?
  - How well the model forecast new events (generalisation)?
  - How superior is one model compared to another?
  - Which is more important, precision or trend?
- Answers:
  - Validation procedures.
  - Metrics or performance measures.
  - Statistical tests.

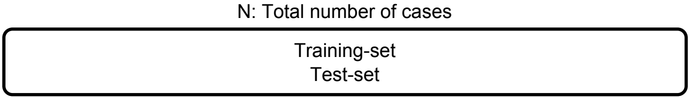
Model prediction (P), observations (O), true state (T)



- a) model with no skill
- b) ideal model
- Reproduced from Stow et al. (2009) and Allen (2009)

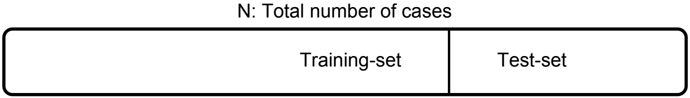
Goodness-of-fit vs generalisation

- Fitting:



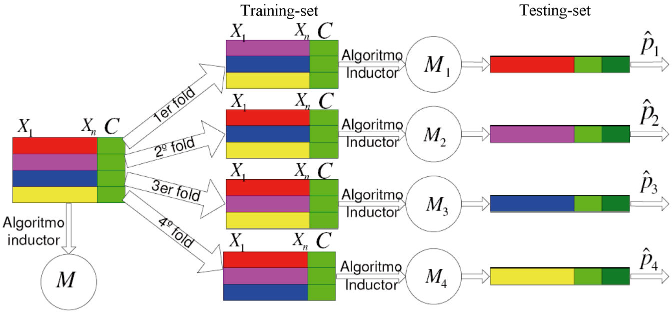
  - Chances of over-fitting.

- Generalization $\rightarrow$ train-test split:



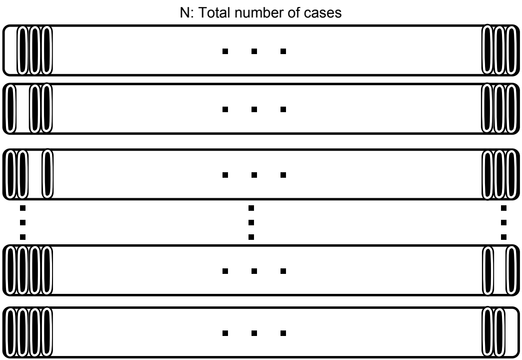
  - Hold-out (commonly 66%-33% split) (Larson, 1931)
  - Hold-out depends on how fortunate the train-test split is.

K-fold cross-validation (CV)



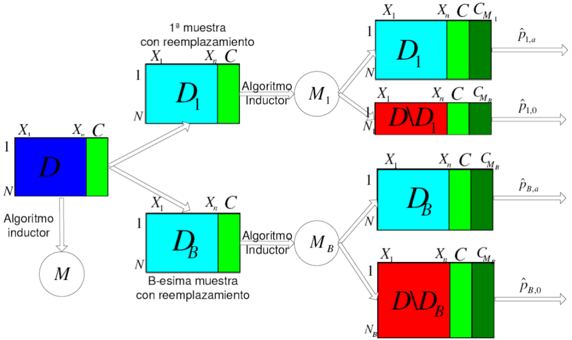
- Performance is the average of k models (Lachenbruch and Mickey, 1968; Stone, 1974).
- All data is eventually used for testing.
- Still sensitive to data split: stratified, repeated (Bouckaert and Frank, 2004).
- Reproduced from Pérez et al. (2005).

Leave-one-out cross-validation (LOOCV)



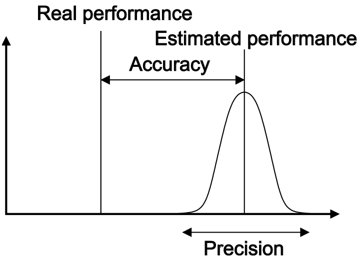
- N models, N-1 cases for training and 1 case for testing (Mosteller and Tukey, 1968).
- Suitable for small datasets, more computationally expensive.
- Variance of the error is the largest, but less biased.
- It can be used for more stable parameters (less variance)

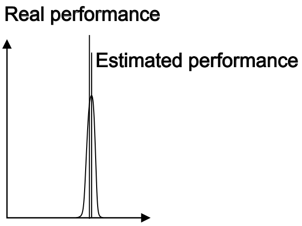
Bootstrapping (0.632 bootstrap)



- A case has a 0.632 probability of being picked for training-set (Efron, 1979).
- error $= 0.632 * e_{test}$ (generalisation) $+ 0.368 * e_{training}$ (fit).
- At least 100 resamplings, some studies suggest 10000.
- Reproduced from Pérez et al. (2005).

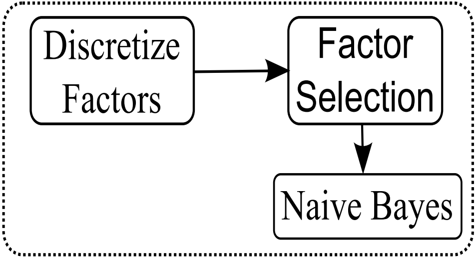
Sumarizing



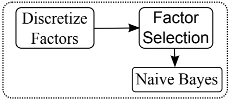


- Increasing data partitions leads to ...
    - more accurate performance estimation (+).
    - more variance in the performance estimation, less precise (-).
    - more computationally expensive (-).
- K-fold cross-validation: trade-off (Rodríguez et al., 2010).

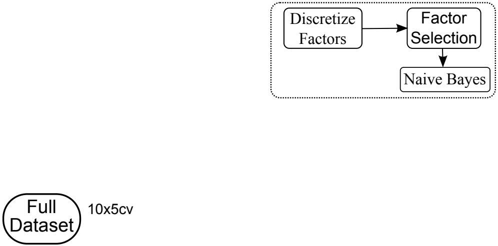
Pipeline validation in filter methods



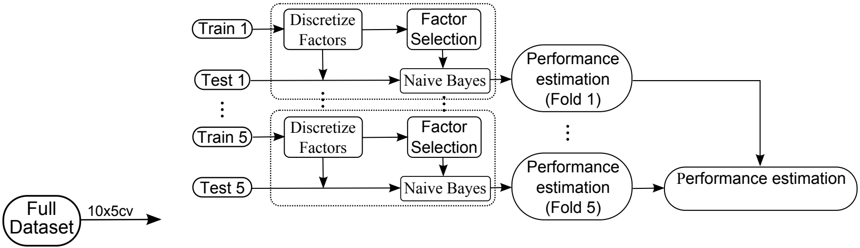
Pipeline validation in filter methods



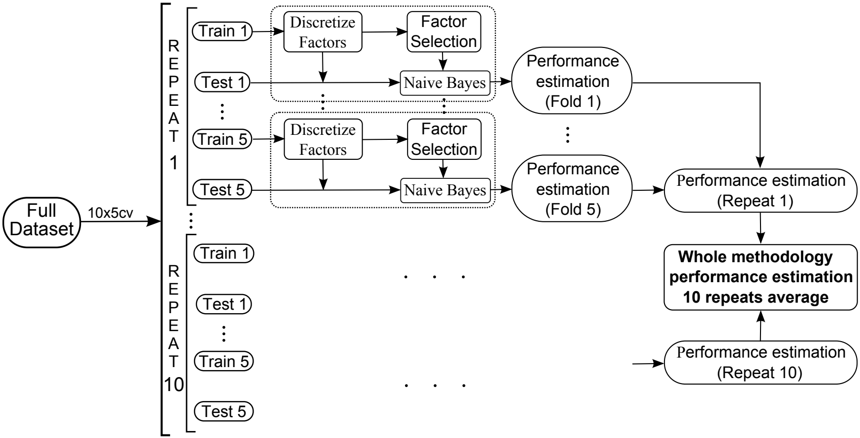
Pipeline validation in filter methods



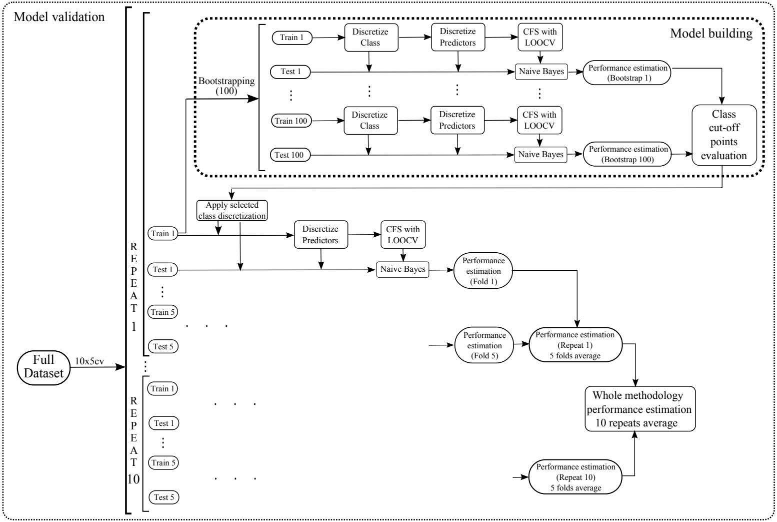
Pipeline validation in filter methods



Pipeline validation in filter methods



Pipeline validation in wrapper methods



Outline

Introduction to metrics

- Each metric shows a different property of the model (Holt et al., 2005; Fernandes et al., 2010)
- Low vs high:
  - Lower is better (error)
  - Higher is better (performance)
- Bounds:
  - Boundless
  - Between 0 and 1
  - Between 0 and 100%

Outline

Numeric prediction metrics

Performance measure	Formula	Performance measure	Formula
mean-squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$	root relative squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
root mean-squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$	relative absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$
mean absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$	correlation coefficient	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i(p_i-\bar{p})(a_i-\bar{a})}{n-1}$,
relative squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$, where $\bar{a}=\frac{1}{n}\sum_i a_i$		$S_p=\frac{\sum_i(p_i-\bar{p})^2}{n-1}$, and $S_A=\frac{\sum_i(a_i-\bar{a})^2}{n-1}$

- Where p are predicted values and a are the actual values.
- Mean-squared error: outliers $\rightarrow$ mean absolute error.
- Relative squared error: relative to the mean of actual values.
- Correlation coeficient: bounded between 1 and -1.
- Reproduced from Witten and Frank (2005).

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum(p-a)^2}{n}}$$

- Goodness of fit between model and observations.
- The closer to 0 the better is the fit.
- If RMSE greater than variance of observations: poor model.
- Reproduced from Allen (2009)

Nash Sutcliffe Model Efficiency)

$$ME = 1 - \frac{\sum_{n=1}^{N}(a_n - p_n)^2}{\sum_{n=1}^{N}(a_n - \bar{a}))^2}$$

- Ratio of the model error to data variability.
- Levels: >0.65 excellent, >0.5 very good, >0.2 good, <0.2 poor Márechal (2004).
- Proposed in Nash and Sutcliffe (1970), reproduced from Allen (2009)

Percentage Model Bias

$$Pbias = \frac{\sum_{n=1}^{N}(a_n - p_n)}{\sum_{n=1}^{N}(a_n)} * 100$$

- Sum of model error normalised by the data.
- Measure of underestimation or overestimation of observations.
- Levels: <10 excellent, <20 very good, <40 good, >40 poor Márechal (2004).
- Reproduced from Allen (2009)

Pearson correlation coefficient (R)

$$R = \frac{\sum_{n=1}^{N}(a_n - \overline{a})(p_n - \overline{p})}{\sqrt{\sum_{n=1}^{N}(a_n - \overline{a})^2 \sum_{n=1}^{N}(p_n - \overline{p})^2}} * 100$$

- Quality of fit of a model to observations.
- R = 0, no relationship.
- R = 1, perfect fit.
- Square of the correlation coefficient (R_2):
- percentage of the variability in data accounted for by the model.
- Reproduced from Allen (2009).

Reliability Index (RI)

$$RI = \exp\sqrt{\frac{1}{n}\sum_{n=1}^{N}(\log\frac{a_n}{p_n})_2}$$

- Factor of divergence between predictions and data.
- RI = 2, means a divergence on average within of a multiplicative factor of 2.
- RI the closer to 1 the better.
- Reproduced from Allen (2009)

Cost functions

- Do all errors have the same weight, cost or implications?
- Scaling of differences between p and a.
- E.g. RMSE scaled by the variance of data (Holt et al., 2005).
- Different cost values depending on the type of error.

Table 5.5 **Default cost matrixes: (a) a two-class case and (b) a three-class case.**

		Predicted class	
		yes	no
Actual class	yes	0	1
	no	1	0

(a)

		Predicted class		
		a	b	c
Actual class	a	0	1	1
	b	1	0	1
	c	1	1	0

(b)

Outline

Confusion matrix: accuracy and true positive

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

- $Accuracy = \frac{TP+TN}{\#cases}$
- $True\ Positive\ Rate = \frac{TP}{TP+FN}$
- Higher is better for both.

Confusion matrix: accuracy and true positive

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

- $Accuracy = \frac{TP+TN}{\#cases}$
- $True\ Positive\ Rate = \frac{TP}{TP+FN}$
- Higher is better for both.

Confusion matrix: accuracy and true positive

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

- $Accuracy = \frac{TP+TN}{\#cases}$
- $True\ Positive\ Rate = \frac{TP}{TP+FN}$
- Higher is better for both.

Confusion matrix: accuracy and true positive

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

- $Accuracy = \frac{TP+TN}{\#cases}$
- $True\ Positive\ Rate = \frac{TP}{TP+FN}$
- Higher is better for both.

Confusion matrix: accuracy and true positive

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

- $Accuracy = \frac{TP+TN}{\#cases}$
- $True\ Positive\ Rate = \frac{TP}{TP+FN}$
- Higher is better for both.

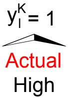
Confusion matrix: accuracy and true positive

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

- $Accuracy = \frac{TP+TN}{\#cases}$
- $True\ Positive\ Rate = \frac{TP}{TP+FN}$
- Higher is better for both.

Brier Score

- (Brier, 1950; van der Gaag et al., 2002; Yeung et al., 2005)
- $Brier\ Score = \frac{1}{\#cases} \sum_{k=1}^{\#cases} \sum_{l=1}^{\#classes} (p_l^k - y_l^k)^2$
- Lower is better (contrary to accuracy & true positive)
- Levels: <0.10 excellent, <20 superior, <0.30 adequate, <0.35 acceptable, >0.35 insuficient (Fernandes, 2011)



Brier Score

- (Brier, 1950; van der Gaag et al., 2002; Yeung et al., 2005)
- $Brier\ Score = \frac{1}{\#cases} \sum_{k=1}^{\#cases} \sum_{l=1}^{\#classes} (p_l^k - y_l^k)^2$
- Lower is better (contrary to accuracy & true positive)
- Levels: <0.10 excellent, <20 superior, <0.30 adequate, <0.35 acceptable, >0.35 insuficient (Fernandes, 2011)

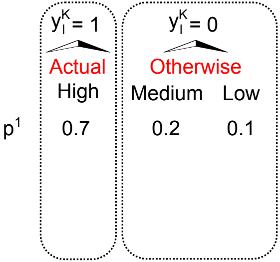
$y_l^K = 1$ $y_l^K = 0$

Actual Otherwise

High Medium Low

Brier Score

- (Brier, 1950; van der Gaag et al., 2002; Yeung et al., 2005)
- $Brier\ Score = \frac{1}{\#cases} \sum_{k=1}^{\#cases} \sum_{l=1}^{\#classes} (p_l^k - y_l^k)^2$
- Lower is better (contrary to accuracy & true positive)
- Levels: <0.10 excellent, <20 superior, <0.30 adequate, <0.35 acceptable, >0.35 insuficient (Fernandes, 2011)



$(0.7-1)^2 + (0.2-0)^2 + (0.1-0)^2 = 0.14$

Brier Score

- (Brier, 1950; van der Gaag et al., 2002; Yeung et al., 2005)
- $\textit{Brier Score} = \frac{1}{\#cases} \sum_{k=1}^{\#cases} \sum_{l=1}^{\#classes} (p_l^k - y_l^k)^2$
- Lower is better (contrary to accuracy & true positive)
- Levels: <0.10 excellent, <20 superior, <0.30 adequate, <0.35 acceptable, >0.35 insuficient (Fernandes, 2011)

	$y_l^K = 1$	$y_l^K = 0$	
	Actual	Otherwise	
	High	Medium	Low
p^1	0.7	0.2	0.1
p^2	0.8	0.1	0.1

$(0.7-1)^2 + (0.2-0)^2 + (0.1-0)^2 = 0.14$

$(0.8-1)^2 + (0.1-0)^2 + (0.1-0)^2 = 0.06$

Brier Score

- (Brier, 1950; van der Gaag et al., 2002; Yeung et al., 2005)
- $Brier\ Score = \frac{1}{\#cases} \sum_{k=1}^{\#cases} \sum_{l=1}^{\#classes} (p_l^k - y_l^k)^2$
- Lower is better (contrary to accuracy & true positive)
- Levels: <0.10 excellent, <20 superior, <0.30 adequate, <0.35 acceptable, >0.35 insuficient (Fernandes, 2011)

	$y_l^K = 1$	$y_l^K = 0$		
	Actual	Otherwise		
	High	Medium	Low	
p^1	0.7	0.2	0.1	$(0.7-1)^2 + (0.2-0)^2 + (0.1-0)^2 = 0.14$
p^2	0.8	0.1	0.1	$(0.8-1)^2 + (0.1-0)^2 + (0.1-0)^2 = 0.06$
p^3	0.1	0.5	0.4	$(0.1-1)^2 + (0.5-0)^2 + (0.4-0)^2 = 1.22$

Brier Score

- (Brier, 1950; van der Gaag et al., 2002; Yeung et al., 2005)
- $Brier\ Score = \frac{1}{\#cases} \sum_{k=1}^{\#cases} \sum_{l=1}^{\#classes} (p_l^k - y_l^k)^2$
- Lower is better (contrary to accuracy & true positive)
- Levels: <0.10 excellent, <20 superior, <0.30 adequate, <0.35 acceptable, >0.35 insuficient (Fernandes, 2011)

	$y_l^K = 1$	$y_l^K = 0$		
	Actual	Otherwise		
	High	Medium	Low	
p^1	0.7	0.2	0.1	$(0.7-1)^2 + (0.2-0)^2 + (0.1-0)^2 = 0.14$
p^2	0.8	0.1	0.1	$(0.8-1)^2 + (0.1-0)^2 + (0.1-0)^2 = 0.06$
p^3	0.1	0.5	0.4	$(0.1-1)^2 + (0.5-0)^2 + (0.4-0)^2 = 1.22$
p^4	0.4	0.5	0.1	$(0.4-1)^2 + (0.5-0)^2 + (0.1-0)^2 = 0.62$

Brier Score

- (Brier, 1950; van der Gaag et al., 2002; Yeung et al., 2005)
- $\textit{Brier Score} = \frac{1}{\#cases} \sum_{k=1}^{\#cases} \sum_{l=1}^{\#classes} (p_l^k - y_l^k)^2$
- Lower is better (contrary to accuracy & true positive)
- Levels: <0.10 excellent, <20 superior, <0.30 adequate, <0.35 acceptable, >0.35 insuficient (Fernandes, 2011)

	$y_l^K = 1$	$y_l^K = 0$		
	Actual	Otherwise		
	High	Medium	Low	
p^1	0.7	0.2	0.1	$(0.7-1)^2 + (0.2-0)^2 + (0.1-0)^2 = 0.14$
p^2	0.8	0.1	0.1	$(0.8-1)^2 + (0.1-0)^2 + (0.1-0)^2 = 0.06$
p^3	0.1	0.5	0.4	$(0.1-1)^2 + (0.5-0)^2 + (0.4-0)^2 = 1.22$
p^4	0.4	0.5	0.1	$(0.4-1)^2 + (0.5-0)^2 + (0.1-0)^2 = 0.62$

Brier Score: (0.14 + 0.06 +1.22 + 0.62) / 4 = 0.51

Normalized Brier Score: 0.51 / 2 = 0.255

Percent Reduction in Error (PRE)

- The relevance of a performance gain.
- A 2% gain of an already highly accurate classifier (90%)
- ... more relevant than with low starting accuracy (50%)
- $$PRE = 100 \cdot \frac{EB - EA}{EB}$$
- EB is the error in the first method (Error Before)
- EA is in the second method (Error After)

Accuracy paradox

$$\mathrm{A}(M) = \frac{TN + TP}{TN + FP + FN + TP}$$ where

TN is the number of true negative cases
FP is the number of false positive cases
FN is the number of false negative cases
TP is the number of true positive cases

	Predicted Negative	Predicted Positive
Negative Cases	9,700	150
Positive Cases	50	100

$$\mathrm{A}(M) = \frac{9,700 + 100}{9,700 + 150 + 50 + 100} = 98.0\%$$

	Predicted Negative	Predicted Positive
Negative Cases	9,850	0
Positive Cases	150	0

$$\mathrm{A}(M) = \frac{9,850 + 0}{9,850 + 150 + 0 + 0} = 98.5\%$$

- Mainly with unbalanced datasets (Zhu and Davidson, 2007; Abma, 2009).
- Reproduced from Wikipedia (2011).

Minimum Description Length (MDL) principle

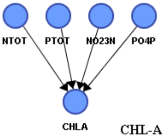
- Kiss rule: Keep It Simple ... Occam's Razor:
- The simplest explanation is the most likely to be true ...
- ... and is more easily accepted by others ...
- ... but, it is not necessarily the truth.

- The more a sequence of data can be compressed, ...
- ... the more regularity has been detected in the data:
- MDL: Minimum Description Length (Rissanen, 1978)

- Trade-off between performance and complexity.
- Is MDL false? Domingos (1999); Grünwald et al. (2005)
- Trade-off between mechanism and robust parameters.
- If two models have same performance then keep the simplest.

Example complex vs simple

EXPERTS SPECIFIED NETWORK



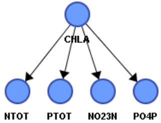
NTOT probability table

Good	Moderate
94,565	5,435

CHL-A probability table

NTOT	PTOT	NO23N	PO4P	Good	Moderate
Good	Good	Good	Good	22,414	77,586
			Moderate	50,000	50,000
		Moderate	Good	0,222	99,778
			Moderate	2,941	97,059
	Moderate	Good	Good	20,186	79,814
			Moderate	1,515	98,485
		Moderate	Good	0,617	99,383
			Moderate	98,485	1,515
Moderate	Good	Good	Good	50,000	50,000
			Moderate	50,000	50,000
		Moderate	Good	50,000	50,000
			Moderate	50,000	50,000
	Moderate	Good	Good	50,000	50,000
			Moderate	97,059	2,941
		Moderate	Good	2,941	97,059
			Moderate	50,000	50,000

NAIVE BAYES CLASSIFIER



NTOT probability table

CHLA	Good	Moderate
Good	97,059	2,941
Moderate	94,000	6,000

CHL-A probability table

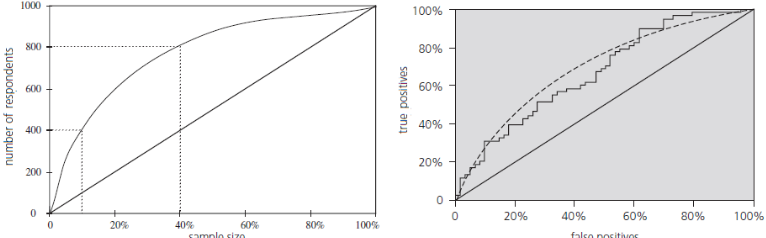
Good	Moderate
18,478	81,522

	Accuracy		True positive		False positive	
	NBC	Expert	NBC	Expert	NBC	Expert
All coastal	82.5±4.1	80.1±5.9	0.18±0.1	0.19±0.08	0.07±0.03	0.08±0
EGOF	84.5±7.8	88.9±11.4	0.5±0	0.5±0	0.06±0	0.06±0
WGOF	73.3±12.6	69.1±13.5	0.13±0	0.14±0.04	0.13±0.05	0.17±0

Lift chart, ROC curve, recall-precision curve

Different measures used to evaluate the false positive versus the false negative tradeoff.

	Domain	Plot	Axes	Explanation of axes
lift chart	marketing	TP vs. subset size	TP	number of true positives
			subset size	$\frac{TP+FP}{TP+FP+TN+FN} \times 100\%$
ROC curve	communications	TP rate vs. FP rate	TP rate	$tp = \frac{TP}{TP+FN} \times 100\%$
			FP rate	$fp = \frac{FP}{FP+TN} \times 100\%$
recall–precision curve	information retrieval	recall vs. precision	recall	same as TP rate tp
			precision	$\frac{TP}{TP+FP} \times 100\%$



Outline

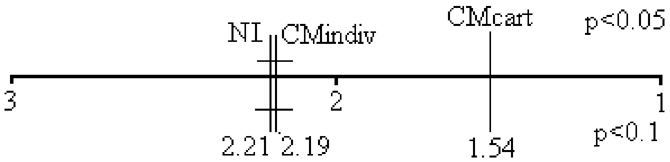
Corrected paired t-test

- Statistical comparisons of the performance.
- Ideal: test over several datasets of size N.
- Null hypothesis that the mean difference is zero. Errors:
- Type I: prob. the test rejects the null hypothesis incorrectly
- Type II: prob. the null hypot. is not rejected with difference.
- Reality: only one dataset of size N to get all estimates.
- Problem: Type I errors exceed the significance level
- Solution: heuristic versions of the *t-test*.

(Nadeau and Bengio, 2003; McCluskey and Lalkhen, 2007; Kotsiantis, 2007; Fernandes, 2011)

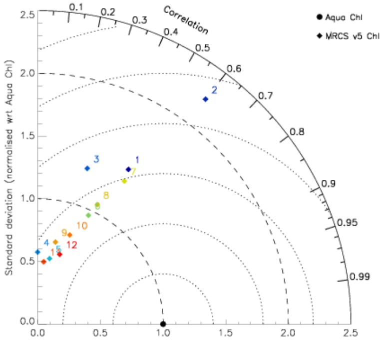
- Comparing MULTIPLE methods over ONE datasets.
- Comparing ONE methods over MULTIPLE datasets.

Critical difference diagrams



- Proposed by Demšar (2006)
- Revised Friedman plus Shaffer's static post-hoc test (García and Herrera, 2008).
- Comparing MULTIPLE methods over MULTIPLE datasets.
- Shows average rank of methods superiority in datasets.
- No significant difference: line connecting methods.
- More datasets: more easy to find significant diferences.

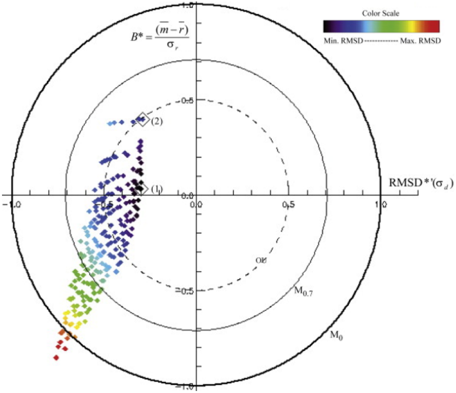
Taylor diagrams



$$E'^2 = \sigma_f^2 + \sigma_r^2 - 2\sigma_f\sigma_r R; c^2 = a^2 + b^2 - 2ab\cos\varphi$$

- Simultaneously: RMS difference, correlation and std. dev.
- R: correlation p & a; E': RMS diff.; σ_f^2 & σ_r^2: variances p & a.
- Proposed in Taylor (2001), reproduced from Allen (2009).

Target diagrams



- RMSE in X-axis; Bias in Y-axis.
- *p* Std. Dev. larger (x>0) than *a*; Bias positive (Y>0) or not.
- Reproduced from Jolliff et al. (2009) and Allen (2009).

Multivariate aproaches

- Uni-variate & multi-variate metrics summarize model skill.
- Multi-variate approaches: simultaneous examination of several variables variation to each other spatially and temporally.

Principal Componet Analysis (PCA) (Jolliffe, 2002).

- Show the relationship between several variables in 2D space.

Multi Dimensional Scalling (MDS) (Borg and Groenen, 2005).

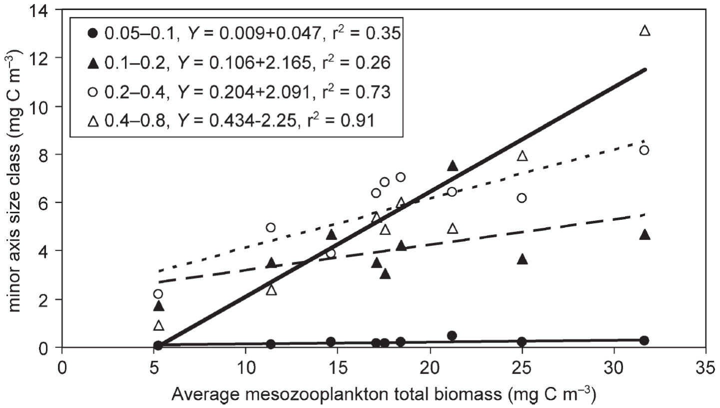
- Exploring similarities or dissimilarities in data

Self organizing Maps (SOM) (Kohonen and Maps, 2001).

- Produce a low-dimensional discretized representation of the observations.

Outline

Zooplankton biomass models



- Several models fits with squared error.
- Reproduced from Irigoien et al. (2009).

An example of anchovy recruitment

Bins	Metrics	Equal frequency	Expert	*Max_mean_tp*	*Max_accuracy*
2	10 × 5cv Acc.	73.7 ± 4.9%	71.2 ± 3.9%	65.1 ± 5.5%	67.4 ± 4.9%
	Best 5cv Acc.	82.1 ± 18.9%	76.8 ± 18.7%	74.3 ± 9%	76.8 ± 16.5%
	Best fold Acc.	100%	100%	87.5%	100%
	Brier score	0.08 ± 0.2	0.08 ± 0.03	0.19 ± 0.05	0.10 ± 0.07
	TP Low	79.9% (<1550; 20)	77.1% (<1500; 21)	61.8% (>1050; 14)	74.4% (<3600; 33)
	TP High	67.4% (>1550; 19)	67.7% (>1500; 17)	54% (>1050; 25)	28.4% (>3600; 6)
3	10 × 5cv Acc.	41.3 ± 9.2%	47.4 ± 7.1%	44.9 ± 5%	47.1 ± 7.6%
	Best 5cv Acc.	53.9 ± 10.5%	55.7 ± 21.5%	51.4 ± 20%	58.9 ± 10.4%
	Best fold Acc.	75%	100%	75%	75%
	Brier score	0.21 ± 0.05	0.16 ± 0.03	0.24 ± 0.05	0.23 ± 0.04
	TP low	47.1% (<1000; 13)	75.6% (<1500; 19)	47.3% (<1200; 16)	50.4% (<1500; 19)
	TP medium	32.9% (1000–2400; 13)	24.3% (1500–3000; 9)	27% (1200–3250; 14)	24.4% (1500–3250; 11
	TP high	51.8% (>2400;13)	28.1% (>3000; 11)	39.4% (>3250; 9)	41% (>3250; 9)
4	10 × 5cv Acc.	33.4 ± 6.3%		30.8 ± 4.1%	26.93 ± 6.8%
	Best 5cv Acc.	41.4 ± 12.5%		36.4 ± 18.1%	38.2 ± 11.7%
	Best fold Acc.	62.5%	–	62.5%	50%
	Brier score	0.25 ± 0.04		0.34 ± 0.06	0.31 ± 0.04
	TP low	49.7% (<850; 10)		36.5% (<1050; 14)	43.3% (<1050; 14)
	TP med. I	10% (850–1550; 10)		10.8% (1050–1900; 9)	11.3% (1050–1900; 9)
	TP med. II	27.7% (1550–3250; 10)		15% (1900–3350; 9)	11.3% (1900–3350; 9)
	TP high	51.8% (>3250; 9)		35.7% (3350>; 8)	30.4% (>3350; 8)

- Performance reported depending on validation schema.
- Reproduced from Fernandes et al. (2010).

Phytoplankton classification

Table III: Output of the significance test

Iteration	RF	TAN
1	90.88	88.95
2	90.7	88.77
3	90	89.12
4	91.05	89.3
5	91.75	88.95
	(v/ /*)	(0/5/0)

The percent of correctly classified instances is compared with the classifications performed using RF and TAN algorithms. Annotation (v/ /*) corresponds to the number of iterations in which the TAN algorithm is significantly better (v), similar () or worse (*) than RF.

Table II: Output of the significance test

Iteration	RF	TAN
1	99.43	99.1
2	99.47	99.15
3	99.52	99.12
4	99.48	99.07
5	99.43	99.09
	(v/ /*)	(0/0/5)

The percentage of correctly classified instances using RF anc algorithms are compared. Annotations (v/ /*) correspond to the nu of iterations in which TAN algorithm is significantly better (v), simila worse (*) than RF.

- Without (Table III) and with (Table II) statistical differences (corrected paired t-test).
- Reproduced from Zarauz et al. (2009) and Zarauz et al. (2008).

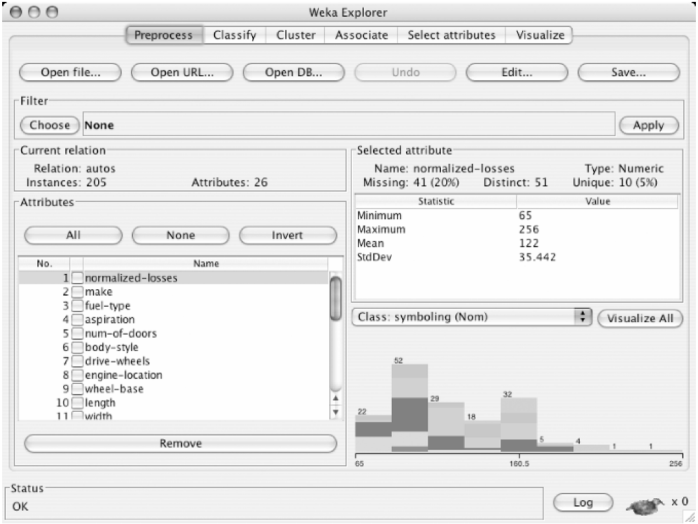
Zooplankton classification

Merger evaluation		DataSet1	DataSet2	DataSet3
Before	Accuracy (%)	64.7	85.7	82
After first 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
	CPU-time	3:01:39	0:32:34	1:30:47
	CPU-time CM	0:17:37	0:16:07	0:17:31
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	–
	CPU-time	1:57:40	0:17:45	–

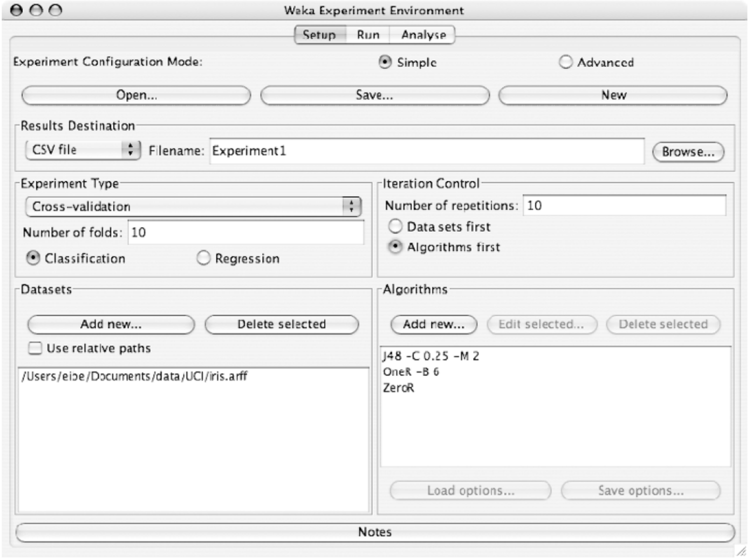
- Reproduced from Fernandes et al. (2009).

Outline

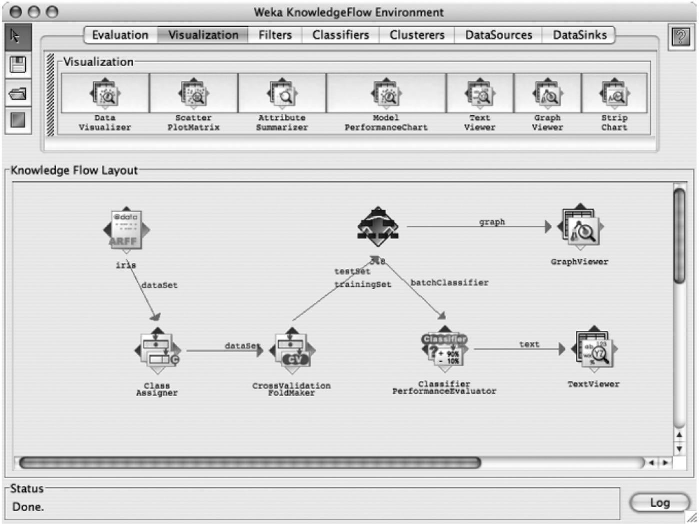
Weka explorer



Weka experimenter



Weka knowledge flow



Outline

Abma, B. (2009). *Evaluation of requirements management tools with support for traceability-based change impact analysis*. PhD thesis, University of Twente, Enschede, The Netherlands.

Allen, J. (2009). D2.7 user guide and report outlining validation methodology. *Deliverable in project Marine Ecosystem Evolution in a Changing Enviroment (MEECE)*.

Borg, I. and Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Verlag.

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

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

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

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

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

Fernandes, J. (2011). *Data analysis advances in marine science for fisheries management: Supervised classification applications*. PhD thesis, University of the Basque Country, San Sebastian, Guipuzkoa, Spain.

Fernandes, J. A., Irigoien, X., Boyra, G., Lozano, J. A., and Inza, I. (2009). 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., and Bode, A. (2010). Fish recruitment prediction, using robust supervised classification methods. *Ecol. Model.*, 221(2):338–352.

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.

García, S. and 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.

Grünwald, P., Myung, I., and Pitt, M. (2005). *Advances in minimum description length: Theory and applications*. The MIT Press.

Holt, J., Allen, J., Proctor, R., and Gilbert, F. (2005). Error quantification of a high-resolution coupled hydrodynamic-ecosystem coastal-ocean model: Part 1 model overview and assessment of the hydrodynamics. *Journal of Marine Systems*, 57(1-2):167–188.

Irigoien, X., Fernandes, J., Grosjean, P., Denis, K., Albaina, A., and 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.

Jolliff, J., Kindle, J., Shulman, I., Penta, B., Friedrichs, M., Helber, R., and Arnone, R. (2009). Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *Journal of Marine Systems*, 76(1-2):64–82.

Jolliffe, I. (2002). Principal component analysis. *Encyclopedia of Statistics in Behavioral Science*.

Kohonen, T. and Maps, S. (2001). Springer series in information sciences. *New York, New York*.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Inform.*, 31:249–268.

Lachenbruch, P. and Mickey, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, pages 1–11.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Educ. Psychol.*, 22(1):45–55.

Márechal, D. (2004). *A soil-based approach to rainfall-runoff modelling in ungauged catchments for England and Wales*. PhD thesis, Cranfield University, Cranfield, UK.

McCluskey, A. and Lalkhen, A. G. (2007). Statistics iv: Interpreting the results of statistical tests. *Continuing Education in Anaesthesia, Critical Care & Pain*, 7(6):208–212.

Mosteller, F. and 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.

Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Mach. Learn.*, 52(3):239–281.

Nash, J. and Sutcliffe, J. (1970). River flow forecasting through conceptual models part i–a discussion of principles. *Journal of hydrology*, 10(3):282–290.

Pérez, A., Larrañaga, P., and I., I. (2005). Estimar, descomponer y comparar el error de mala clasificación. In *Primer Congreso Español de Informática*.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*, 14:465–471.

Rodríguez, J. D., Pérez, A., and 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.

Schirripa, M. J. and 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.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statistical Society, Series B*, 36.

Stow, C., Jolliff, J., McGillicuddy Jr, D., Doney, S., Allen, J., Friedrichs, M., and Rose, K. (2009). Skill assessment for coupled biological/physical models of marine systems. *Journal of Marine Systems*, 76(1-2):4–15.

Taylor, K. (2001). Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res*, 106(D7):7183–7192.

van der Gaag, L. C., Renooij, S., Witteman, C. L. M., Aleman, B. M. P., and Taal, B. G. (2002). Probabilities for a probabilistic network: a case study in oesophageal cancer. *Artif. Intell. Med.*, 25(2):123–148.

Wikipedia (2011). Accuracy paradox. [Online; accessed 15-September-2011].

Witten, I. H. and Frank, E. (2005). *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, CA, USA*.

Yeung, K. Y., Bumgarner, R. E., and 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.

Zarauz, L., Irigoien, X., and 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., and 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.

Zhu, X. and Davidson, I. (2007). *Knowledge discovery and data mining: challenges and realities*. Igi Global.

EURO-BASIN Training Workshop on Introduction to statistical modelling tools, for habitat models development: Model Validation Performance measures Models comparison WEKA: open source software for data mining

Jose A. Fernandes

University of East Anglia / CEFAS
AZTI-Tecnalia
Intelligent Systems Group (University of Basque Country)