

Minería de Datos

II. Experiencias Exitosas en el ISG (UPV/EHU)

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid

Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco

MMCC, 2006-2007

Índice

- 1 **Introducción**
- 2 **Medicina**
- 3 **Bioinformática**
- 4 **Empresa**
- 5 **Industria**
- 6 **Informática**
- 7 **Conclusiones**

Índice

1 Introducción

2 Medicina

3 Bioinformática

4 Empresa

5 Industria

6 Informática

7 Conclusiones

Éxito en minería de datos

- **Ganar dinero**
- **Vivir mejor**: mas seguros, menos riesgos, evitar enfermedades
- **Generar conocimiento** desconocido, avance científico
- **Implantar el sistema** en una organización

Índice

1 Introducción

2 Medicina

3 Bioinformática

4 Empresa

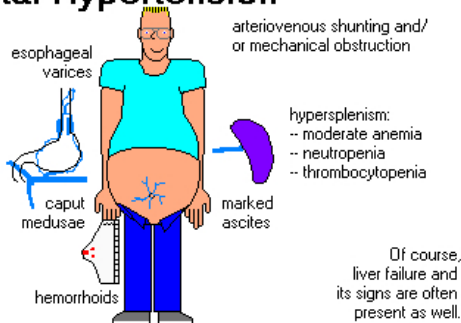
5 Industria

6 Informática

7 Conclusiones

Predicción de la supervivencia en cirróticos tratados con TIPS

Portal Hypertension



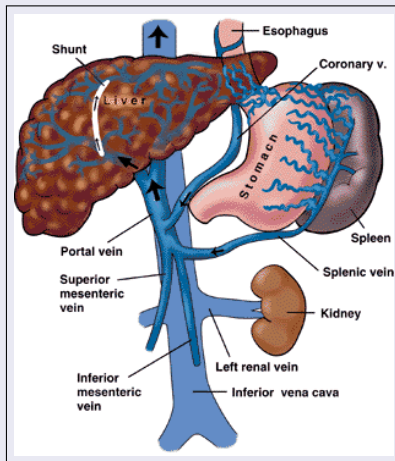
R. Blanco, I. Inza, M. Merino, J. Quiroga, P. Larrañaga (2005) *Journal of Biomedical Informatics*, 38, 376-388

Predicción de la supervivencia en cirróticos tratados con TIPS

Problemática

- **Enfermos crónicos de hígado:** hipertensión portal
- Consecuencia de la hipertensión portal es la **hemorragia de las varices gastroesofágicas** (causa de mortalidad)
- **Desvío portosistémico intrahepático transyugular** (TIPS)
- **Poco conocimiento sobre los efectos** del TIPS en la supervivencia

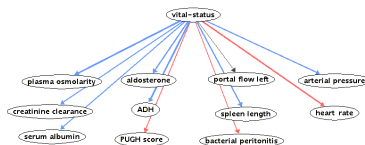
Desvío portosistémico intrahepático transyugular (TIPS)



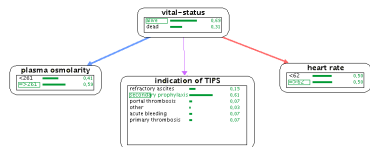
Características de la muestra de pacientes

- Clínica Universitaria de Navarra: 107 pacientes desde Mayo 1991 a Septiembre 1998
- Diagnóstico basado en histología del hígado
- 6 meses periodo crítico considerado por coincidir con el tiempo medio de espera de transplante de hígado
- 77 variables:
 - Historia clínica
 - Laboratorio
 - Ecografía doppler
 - Endoscopia
 - Parámetros hemodinámicos
 - Angiografía

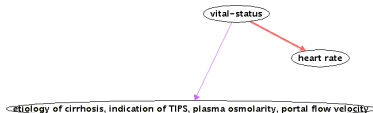
Estructura de varias de las redes Bayesianas



(a) Filter approaches



(b) SeINB_{ws}



(c) SemiNB_{ws}

Resultados con clasificadores Bayesianos

		Porcentaje	No. Variables
Naïve Bayes		88.78 ± 3.06	77
Naïve Bayes	NB_{ws}^g	90.65 ± 2.92	4
Naïve Bayes	NB_{ws}^{eda}	88.78 ± 3.06	27
Seminaïve Bayes	SNB_{ws}^g	88.78 ± 3.06	6
Seminaïve Bayes	SNB_{ws}^{eda}	91.58 ± 2.69	7
TAN		88.78 ± 3.06	77
TAN	TAN_{ws}^g	91.58 ± 2.69	6
TAN	TAN_{ws}^{eda}	91.73 ± 7.08	77
Naïve Bayes	NB_{fs}	93.45 ± 2.40	11
Seminaïve Bayes	SNB_{fs}	92.52 ± 2.55	11
TAN	TAN_{fs}	93.45 ± 2.40	11

Conclusiones

- **Generado conocimiento** acerca de la enfermedad a partir de los datos
- La **selección de variables**:
 - Reduce gastos
 - Evita sufrimiento
 - Agiliza el proceso
 - Permite construir modelos con mayor poder clasificatorio

Índice

- 1 Introducción
- 2 Medicina
- 3 Bioinformática**
- 4 Empresa
- 5 Industria
- 6 Informática
- 7 Conclusiones

A la búsqueda de marcadores genéticos

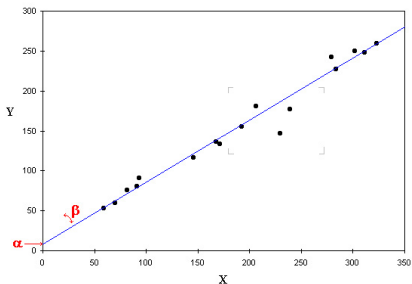
Minería de datos provenientes de microarrays



R. Armañanzas, B. Calvo, I. Inza, P. Larrañaga, I. Bernales, A. Fullaondo, A. M. Zubiaga (2006) *Lecture Notes in Mathematics in Industry*, Springer, en prensa

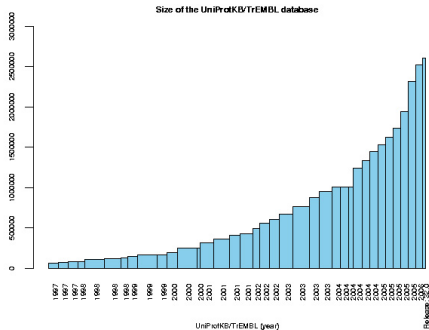
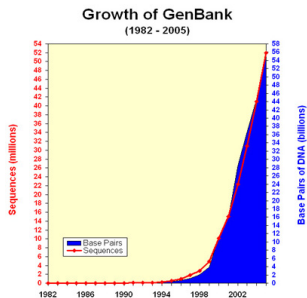
Evolución en Biología: de lo artesanal a

Investigación "hypothesis driven"



Crecimiento bases de datos en genómica y proteómica

GenBank y SwissProt



Bioinformática y Biología Computacional

Bioinformática

Se encarga del **almacenamiento** digital y la **manipulación** de la información biológica

Biología Computacional

Uso de **técnicas computacionales** para generar nuevo conocimiento sobre los sistemas biológicos

Lupus Eritematoso Sistémico

LES (Biett, 1833)

- Origen **desconocido**
- Enfermedad **inflamatoria**
- Características **autoinmunes**
- Puede afectar a **múltiples órganos**
- Semejanza entre las **erupciones cutáneas** (síntoma de la enfermedad) y la **mordedura de un lobo**

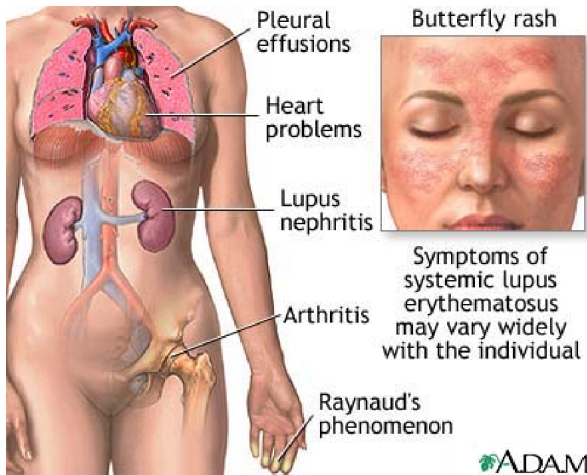
Lupus Eritematoso Sistémico



Lupus Eritematoso Sistémico



Lupus Eritematoso Sistémico



Síndrome Antifosfolípido

SAF (Hughes, 1983)

- Carácter **autoinmune**
- Aparición de **trombosis de repetición**
- Alto número de **abortos esporádicos**
- Alta tendencia a la **coagulación de la sangre**
- **SAF y LES** enfermedades **relacionadas**

Búsqueda de genes relacionados con LES

Objetivos del estudio

- Encontrar **genes involucrados en la enfermedad**
- **A partir de los niveles de expresión génica** que presentan individuos enfermos de LES o SAF e individuos sanos
- Generación de **nuevo conocimiento** acerca de la enfermedad

Diseño del experimento

De muestras de sangre a microarrays

- Muestras de sangre de **12 personas** (Hospital de Basurto y Hospital Universitario Marqués de Valdecilla): 4 LES, 2 SAF y 6 sanas (una no válida)
- **Microarrays de ADN** para medir el **nivel de expresión génica en 22067 genes**. Affymetrix

Problema de clasificación supervisada

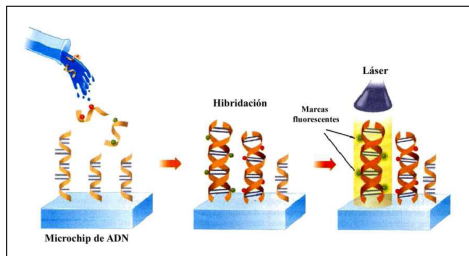
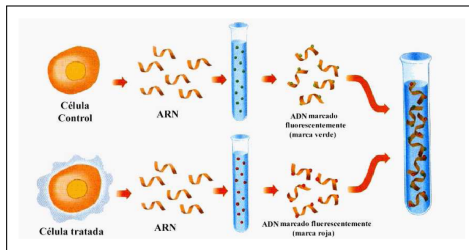
- **40 casos**: 20 LES-sanas, 10 SAF-sanas, 10 sanas-sanas
- 22067 genes reducidos a **8808 genes** (variables)
- **Variable clase**: tres valores (**LES, SAF, sana**)

Biología computacional

Microarrays de ADN

- Bioinformática **desarrollo espectacular**
- Microchips de ADN **transformación de la biología**
- Posible medir el **nivel de expresión de miles de genes** al mismo tiempo
- **Análisis de la información** recogida para la **comprensión** de aspectos involucrados en los **procesos biológicos**

Realización de microarrays (cDNA)



Selección de genes basada en consenso

Identificar genes “prototipo”

- **Discretizaciones:** igual anchura, igual frecuencia, entropía
- **Selección de atributos:** atributos altamente correlados con la clase y con bajo grado de redundancia entre ellos
- **Genes “prototipo”(8 genes):** resultantes de la intersección de las selecciones sobre las distintas bases de datos discretizadas

Buscar genes con patrón similar a los prototipos (150 genes)

- En base a la literatura identificación de genes reguladores de los anteriores (299 genes en total)
- Seleccionar aquellos (19 genes) que regulen a mas de dos genes

Selección de genes basada en consenso

Discretización

De los 169 genes (150 +19), 56 son configurados como constantes al usar la técnica de discretización de la entropía

Problema de clasificación supervisada

- 40 casos
- 113 variables predictoras
- Variable **clase: tres valores** (LES, SAF, sana)

Estimación de la probabilidad de acierto

Naïve Bayes, selective naïve Bayes, TAN, k DB

- **Entorno de desarrollo** de clasificadores Bayesianos:
<http://leo.ugr.es/~elvira>
- Paradigmas de clasificación supervisada: **naïve Bayes, selective naïve Bayes, TAN, k DB** ($k = 2, 3, 4$)
- Estimación de la probabilidad de bien clasificados: *leave one out cross-validation*
- Resultados: **100.0 %** de precisión con naïve Bayes, TAN, k DB ($k = 2, 3, 4$) y **90 %** de precisión con selective naïve Bayes (3 genes)

Descubriendo redes de correulación por medio de bootstrap no-paramétrico (Efron, 1979)

-
1. Repetir B veces
 - 1.1. Muestrear aleatoriamente y con reemplazo N casos de la base de datos original
 - 1.2. Aplicar el algoritmo de inducción a la nueva base de datos muestreada, obteniendo un modelo clasificadorio
 2. Para cada uno de los B modelos inducidos, estimar los niveles de confianza de las características a estudio como la frecuencia relativa de su presencia o ausencia
-

Robustez de los arcos

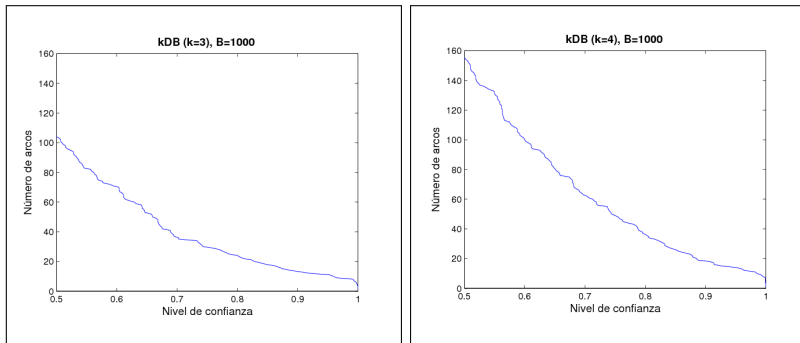


Figura: Número de arcos y niveles de confianza superiores al 50 %. Bootstrap ($B = 1000$) sobre kDB ($k = 3, 4$)

Robustez de los arcos

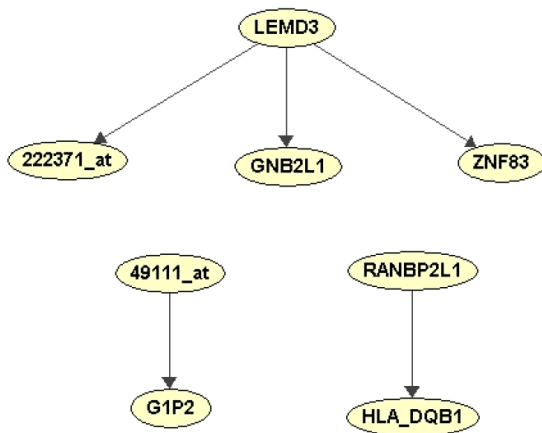


Figura: Estructuras sencillas de dependencias encontradas mediante bootstrap

Robustez de los arcos

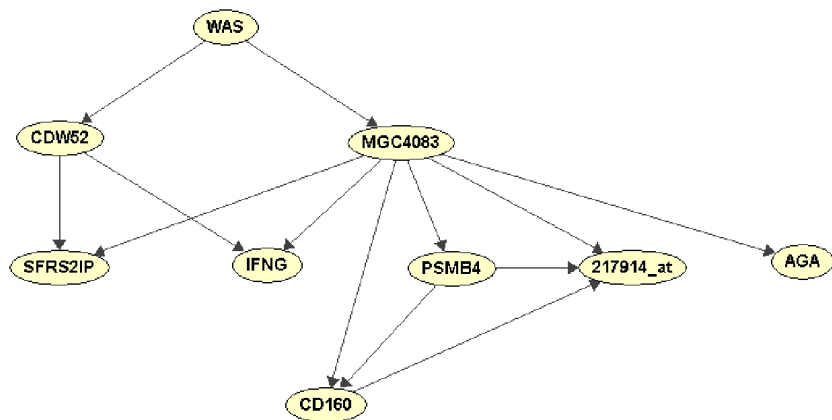


Figura: Estructura de dependencias compleja encontrada mediante bootstrap

Conclusiones

Clasificadores Bayesianos

- Sencillez en la inducción de los modelos
- Gran nivel de transparencia
- Ayuda para descubrir el conocimiento

Bioinformática

- Dominio estimulante para la minería de datos
- Grandes volúmenes de datos
- Problemas novedosos
- Conocimiento descubierto de utilidad

Índice

1 Introducción

2 Medicina

3 Bioinformática

4 Empresa

5 Industria

6 Informática

7 Conclusiones

Predicción de bancarrota en empresas

Con el agua al cuello



S. Dizdarevich, P. Larrañaga, B. Sierra, J. A. Lozano, J. M. Peña (2005). Combining statistical and machine learning based classifiers in the prediction of corporate failure. *Artificial Intelligence in Accounting and Auditing. Volume 6. International Perspective*, Markus Wiener Publishers, 177-211

Predicción de bancarrota en empresas

Fracaso empresarial

- Predicción del fracaso empresarial **estudiado durante los últimos 40 años**
- Importante para **inversores, auditores, acreedores, accionistas, Gobierno, ...**
- **Datos de los informes financieros de la empresa** analizados con técnicas de minería de datos para evaluar y predecir el futuro del estado financiero

Predicción de bancarrota en empresas

Síntomas previos

- **Económicos:**
 - Rentabilidad del capital invertido por debajo de sus costos de oportunidad
 - Ingresos comienzan a ser mas bajos que los gastos
 - Primeros resultados negativos
- **Financieros:**
 - Si el fracaso económico no se corrige, la empresa va a dejar de tener capital suficiente para hacer frente a los pagos
 - Deterioro: deudas mayores que las posesiones (situación negativa del patrimonio neto)

Predicción de bancarrota en empresas

Causas y síntomas

- **Causas:** fracaso administrativo, deficiencia de los sistemas de contabilidad, incapacidad para adaptarse a los cambios del entorno, acometer grandes proyectos, abuso de la financiación a través de deudas, riesgo actual del mundo empresarial,
- **Detección de síntomas:**
 - **Sentido común:** poniendo atención a la realidad cotidiana de la empresa y su entorno: cambio de auditor, renuncia repentina de miembros de dirección, líneas de crédito reducidas o canceladas, exceso de stock,
 - **Análisis del estado contable:** estadística y minería de datos aplicadas a ratios contables homogenizados por sectores y tamaño

Predicción de bancarrota en empresas

Análisis estadísticos

- **Univariante:** Beaver (1966)
- **Multivariante:**
 - **Análisis discriminante:** Altman (1968), Deakin (1972), Edmister (1972), Blum (1974), Libby (1975), Scott (1981), Taffler (1982)
 - **Regresión logística:** Ohlson (1980), Mensah (1983), Zavgren (1985), Casey y Baztchak (1985), Peel y Peel (1987)

Predicción de bancarrota en empresas

Hipótesis y objetivo

- **Hipótesis:** el patrón de información contable de las empresas sanas y de las empresas que fracasan difiere
- **Objetivo:** crear modelos computacionales capaces de predecir con 1, 2, o 3 años de anterioridad el fracaso empresarial

Predicción de bancarrota en empresas

Características de la muestra de empresas

- **Concepto de fracaso** equivale a suspensión de pagos (objetivo y se sabe la fecha)
- **120 compañías**: 60 sanas ($C = s$) y 60 que fracasan ($C = f$) escogidas en 10 provincias españolas de manera pareada (tamaño y sector)
- Información recogida en el **Boletín Oficial del Registro Mercantil Provincial** en 18 meses (Enero 1993 a Julio 1994) época de crisis empresarial en España
- **Datos económicos y financieros** de los tres últimos años

Predicción de bancarrota en empresas

Variables predictoras

- **50 variables** del certificado del balance de comprobación, de la contabilidad de beneficios y pérdidas y del balance financiero
- **Criterios de selección de las variables:** citadas frecuentemente en la literatura del análisis financiero y fáciles de calcular directamente
- **Reducción a 9 variables** por medio de un análisis en componentes principales:
 - X_1 : Activo circulante / Pasivo circulante
 - X_2 : Activo circulante / Activo total
 - X_3 : Resultado neto / Activo total
 - X_4 : Ganancia antes de intereses e impuestos / Costes financieros
 - X_5 : Fondos propios / Deuda total
 - X_6 : Ventas / Fondos propios
 - X_7 : Stocks / Ventas
 - X_8 : Deudores / Ventas
 - X_9 : Cash flow operativo / Activo total
 - C: Saneada (s); fracaso (f)

Predicción de bancarrota en empresas

Ánalysis Discriminante

- $M_3: C = f \Leftrightarrow -2,01 + 2,24X_4 + 2,52X_5 > 0$
- $M_2: C = f \Leftrightarrow -1,46 + 2,36X_2 - 8,15X_3 + 3,13X_5 > 0$
- $M_1: C = f \Leftrightarrow -0,57 + 9,36X_3 + 0,55X_5 > 0$

Predicción de bancarrota en empresas

Regresión Logística

$$M_1 : C = f \Leftrightarrow \frac{e^{-1,28+26,13X_3+1,35X_5}}{1 + e^{-1,28+26,13X_3+1,35X_5}} > 0,5$$

Predicción de bancarrota en empresas

Inducción de Reglas

M_1 :

If ($X_4 < 0,80$) then $C = s$

else if (($X_1 > 1,41$) and ($X_4 > 0,87$) and ($X_7 < 0,36$)) then
 $C = f$

else if (($X_4 > 0,95$) and ($X_5 < 0,22$)) then $C = s$

else if (($X_5 < 0,36$) and ($X_8 > 0,03$)) then $C = f$

else if (($X_4 < 1,90$) and ($X_6 > 4,69$)) then $C = s$

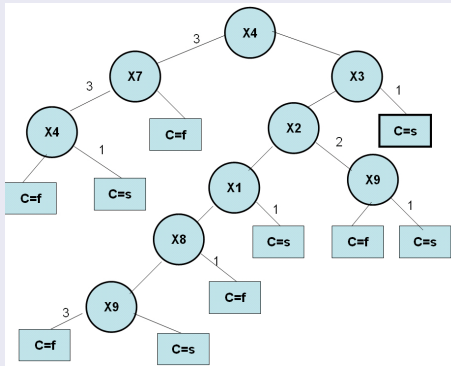
else if (($X_1 > 0,81$) and ($X_1 < 1,38$) and $X_7 < 0,23$) then $C = f$

else if (($X_8 > 0,17$) and ($X_7 < 0,67$)) then $C = s$

else $C = f$

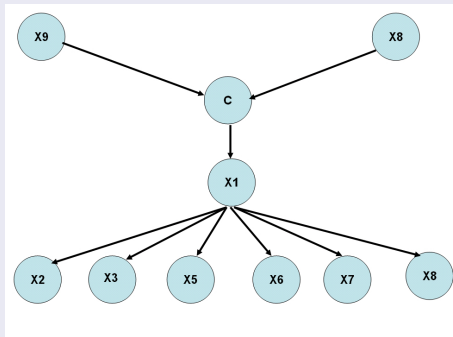
Predicción de bancarrota en empresas

Árbol de Clasificación



Modelo de árbol de clasificación para la predicción un año antes de la bancarrota

Red Bayesiana



Estructura de red Bayesiana para predecir el fracaso empresarial con tres años de antelación

Predicción de bancarrota en empresas

Años antes	AD	RL	IR	AC	RB	Voto
1	78.33	82.50	79.17	80.00	60.83	88.33
2	69.17	69.16	60.00	66.66	62.00	79.13
3	55.00	55.00	45.00	57.50	60.83	73.33

Porcentaje de empresas correctamente clasificadas. Método de validación: 5 *k*-fold crossvalidation

Índice

1 Introducción

2 Medicina

3 Bioinformática

4 Empresa

5 Industria

6 Informática

7 Conclusiones

Seguridad en aeronáutica



Photo Copyright © Michael Engenschwiler

AIRLINERS.NET

Seguridad en aeronáutica

Rebabas en agujeros taladrados



Agujeros taladrados pueden presentar **rebabas**

Seguridad en aeronáutica

Automatizar la detección de rebaba

- **Rebaba**: porción de materia sobrante que forma resalto en los bordes o en la superficie de un objeto manufacturado cualquiera
- Rebabas **no admisibles** desde el punto de vista de la seguridad aeronáutica
- En un **avión gran número de agujeros taladrados**
- Interesa una herramienta que pueda **automatizar** el proceso de **detección de rebaba**
- Detectar las rebabas en agujeros taladrados en un material como el aluminio a través de la **lectura de la señal de un sensor**

Seguridad en aeronáutica

Variables predictoras

- **13 variables predictoras** de dos tipos: variables de configuración (8 variables) y variables de sensor (5 variables)
- **Variables de configuración**: marcan las diversas condiciones en las que se realiza el ensayo
- **Variables sensor** son el resultado obtenido a través de un acelerómetro durante el proceso de taladrado
- El objetivo sería alcanzar un **100 por cien de detección o al menos no tener ningún falso negativo** (agujero dado por bueno y que tenga rebaba superior a 127)

Seguridad en aeronáutica

Weka software

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Undo Edit... Save...

Filter: Choose **None** Apply

Current relation:
Relation: rebaba-weka.filters.unsupervised.attribute.Discretize-F810-M-1.0-R9-13
Instances: 105 Attributes: 14

Attributes: All None Invert

No.	Name
1	BRO
2	NUM
3	YC
4	TRM
5	AV1
6	AV2
7	REC
8	ESP
9	MIN
10	MAX
11	ANG
12	ALT
13	ANC
14	class

Remove

Selected attribute:
Name: class
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

Label	Count
0	69
1	36

Class: class (Nom) Visualize All

Label	Count
0	69
1	36

Seguridad en aeronáutica

Weka software

The screenshot shows the Weka Explorer application window. The 'Classify' tab is selected. The 'Classifier' dropdown is set to 'J48 -C 0.25 -M 2'. Under 'Test options', 'Cross-validation' is selected with 'Folds' set to 10. The '(Nom) class' dropdown is set to 'class'. The 'Start' button is visible. The 'Result list (right-click for options)' shows '10:14:36 - trees.J48' selected. The 'Classifier output' pane displays the following information:

Relation: rebaba-weka.filters.unsupervised.attribute.Discretize-F-B10-N-1.0-R9-13
Instances: 105
Attributes: 14
BR0
NUM
VC
TEM
AV1
AV2
REC
ESP
MIN
MAX
ANG
ALT
AMC
class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

VC = 100: 1 (26.0)
VC = 125
| NUM <= 77: 1 (2.0)
| NUM > 77: 0 (3.0)
VC = 150: 0 (52.0/8.0)
VC = 200: 0 (6.0)
VC = 250: 0 (16.0)

Number of Leaves : 6

Status: OK

Log x0

Seguridad en aeronáutica

Weka software

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier: Choose **J48 -C 0.25 -M 2**

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds
- ☐ Percentage split %

(Nom) class

Result list (right-click for options)

10:14:36 - trees.j48

Classifier output:

```

Number of Leaves :    6
Size of the tree :    8

Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      94           89.5238 %
Incorrectly Classified Instances    11           10.4762 %
Kappa statistic                    0.7527
Mean absolute error                 0.1583
Root mean squared error             0.2974
Relative absolute error             35.0402 %
Root relative squared error         62.6122 %
Total Number of Instances          105


=== Detailed Accuracy By Class ===

TP Rate    FP Rate    Precision    Recall    F-Measure    Class
0.986      0.278      0.872      0.986      0.925        0
0.722      0.014      0.963      0.722      0.825        1

=== Confusion Matrix ===

 a b  <-- classified as
68 1 | a = 0
10 26 | b = 1

```

Status: OK  x 0

Índice

1 Introducción

2 Medicina

3 Bioinformática

4 Empresa

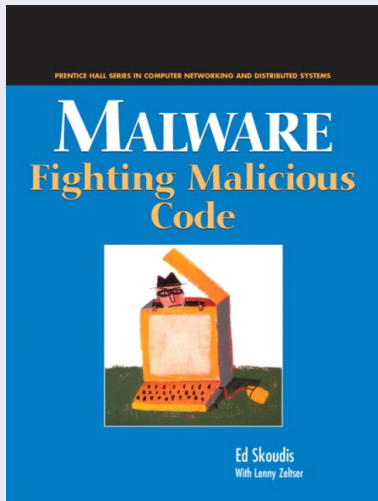
5 Industria

6 Informática

7 Conclusiones

Seguridad informática

Código malicioso



Seguridad informática

Código malicioso

- Fichero caracterizado por **mas de 90.000 variables**
- **Dos categorias:** malicioso vs no malicioso

Retos

- Trabajo **casi a ciegas** por cuestiones de confienciabilidad de los datos
- Clases **no balanceadas**
- **Selección de variables**

Índice

1 Introducción

2 Medicina

3 Bioinformática

4 Empresa

5 Industria

6 Informática

7 Conclusiones

A modo de resumen

- Muchas experiencias exitosas no pueden ser detalladas por cuestiones de **confidenciabilidad**
- Minería de datos metodología aplicable a una **gran variedad de situaciones y problemáticas**
- Para obtener una solución satisfactoria muchas veces es necesario **adaptar los métodos standard a las características del problema**

Minería de Datos

II. Experiencias Exitosas en el ISG (UPV/EHU)

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid

Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco

MMCC, 2006-2007