

Breve Introducción al Aprendizaje Automático con WEKA

Procesamiento del Lenguaje Natural

José María Gómez Hidalgo
<http://www.esp.uem.es/~jimgomez/>

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Índice

- Referencias
- Motivación
- Conceptos básicos
- El proceso de minería de datos
- Selección de atributos
- Algoritmos de aprendizaje
 - PRISM: Inducción de reglas
 - Bayes Ingenuo
 - ID3: árboles de decisión
- Evaluación y visualización

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

Referencias

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Referencias

- Usamos básicamente
 - I. WITTEN, E. FRANK, *Data Mining: Practical Machine Learning Tools and Techniques with Java Applications*, Morgan Kaufmann Publishers, 1999 - 2005
 - QA76.9 .D3 W58 - QA76.9 .D343 W58 Bib. UEM
 - Capítulos 1, 4 y 5
 - WEKA: <http://www.cs.waikato.ac.nz/~ml/weka/>
 - Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *From Data Mining to Knowledge Discovery in Databases*. AI Magazine 17(3), 37-54
<http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

Motivación

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Motivación

- **Objetivo**
 - (semi) automatización de múltiples tareas
 - Predicción de enfermedades
 - Identificación de mareas negras
 - Prevención de fraude financiero
 - Determinación del periodo fértil del ganado vacuno
 - Detección del correo basura o Spam
 - Análisis de tendencias en mercados financieros
 - Etc. Hasta el infinito

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Motivación

- Usualmente, dichas tareas realizadas por experto humano
- Para automatizar
 - Extraer su conocimiento (experiencia) y codificarlo (posiblemente) como reglas
 - Desarrollo de un sistema experto o sistema basado en conocimiento
 - Tarea del ingeniero del conocimiento

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Motivación

- E.g. Recomendación de lentes de contacto
 - En función de
 - Edad (Age), prescripción ocular o enfermedad (Spectacle prescription), astigmatismo (Astigmatism), tasa de lágrimas (Tear production rate)
 - Recomendar a un paciente
 - Lentes blandas (Soft), duras (Hard) o ninguna (None)
 - El experto puede sugerir la regla
 - Si la tasa de lágrimas es baja entonces (recomendar) ninguna

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Motivación

- Problemas
 - Cuello de botella de adquisición del conocimiento
 - Conocimiento difícil de formalizar
 - Expertos no cooperativos
 - Carencia de portabilidad y escalabilidad
 - Se prescinde temporalmente del experto durante la adquisición

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Motivación

- Alternativa
 - Adquirir el conocimiento de manera automática a partir de ejemplos
- Aprendizaje Automático
 - “sistemas que aprenden a cambiar su comportamiento de modo que resulten más efectivos en el futuro”

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Motivación

- (Algunas) ventajas
 - Proceso de adquisición automático
 - Podemos prescindir del experto, y quedarnos con sus datos
 - La tecnología es portable = aprender sobre datos distintos => aplicar sobre dominios nuevos
 - La tecnología es (generalmente) escalable = de hecho, cuantos más (y mejores) datos, mejor funcionará
 - Posibilidad de explotar la actual abundancia de datos
 - *IMPORTANTE: múltiples tareas de PLN se resuelven así de manera (relativamente) sencilla, y existen muchos datos*
 - *IMPORTANTE: software disponible - WEKA*

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Motivación

- (Algunas) desventajas
 - No siempre se alcanza la efectividad del experto
 - El proceso general es bastante más sofisticado
 - Selección de fuentes, recopilación de datos, selección de los más adecuados, estructuración y representación, aprendizaje, comprensión de resultados
 - Descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD)
 - Los datos son confusos, erróneos, incompletos, pocos, con ruido, etc.
 - Muchas técnicas disponibles => comparar

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

Conceptos básicos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Conceptos básicos

- **Terminología**
 - Datos de entrada = ejemplos, ejemplares, instancias = colección de entrenamiento
 - Caracterizados por atributos o rasgos
 - Proceso = entrenamiento o aprendizaje
 - Salida = clasificador
 - Capaz de clasificar nuevos ejemplares (de prueba u operativos)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Conceptos básicos

- E.g. Datos para recomendación de lentes

Age	Spectacle prescription	Astigmatism	Tear production rate	Contact lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	hypermetropo	yes	reduced	none
young	hypermetropo	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	hypermetropo	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	hypermetropo	yes	reduced	none
presbyopic	hypermetropo	yes	normal	none
...

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Conceptos básicos

- E.g. Idem en formato ARFF (WEKA) – Attribute Relation File Format

```
@relation contact-lenses
```

```
@attribute age {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip {myope, hypermetropo}
@attribute astigmatism {no, yes}
@attribute tear-prod-rate {reduced, normal}
@attribute contact-lenses {soft, hard, none}
```

```
@data
```

```
young,myope,no,reduced,none
```

```
young,myope,no,normal,soft
```

```
young,myope,yes,reduced,none
```

```
...
```

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

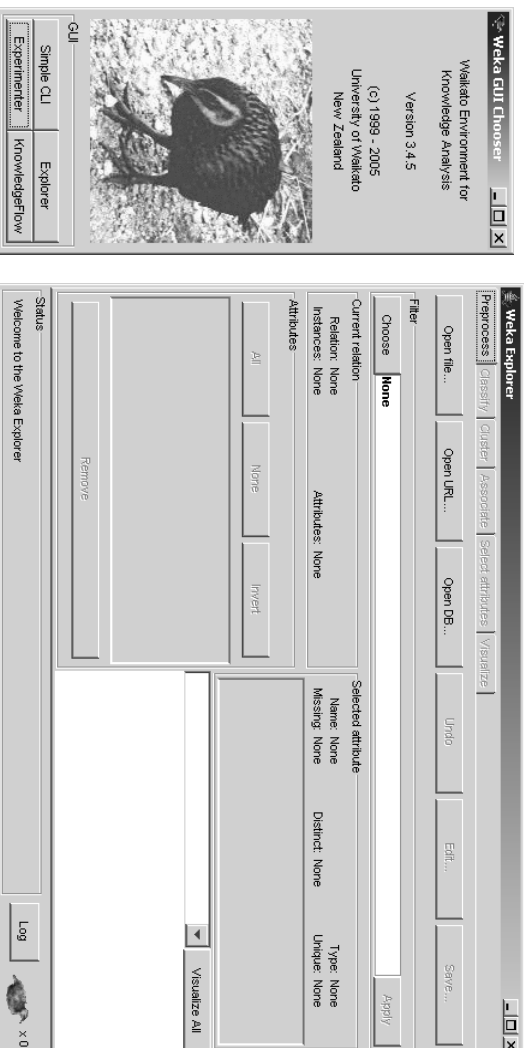
Conceptos básicos

- E.g. Clasificador generado por PRISM
 - Sistema de (9) reglas de clasificación, incluyendo
 - IF astigmatism = no
and tear-prod-rate = normal
and spectacle-prescrip = hypermetrope THEN soft
 - IF astigmatism = yes
and tear-prod-rate = normal
and spectacle-prescrip = myope THEN hard
 - IF tear-prod-rate = reduced THEN none

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Conceptos básicos

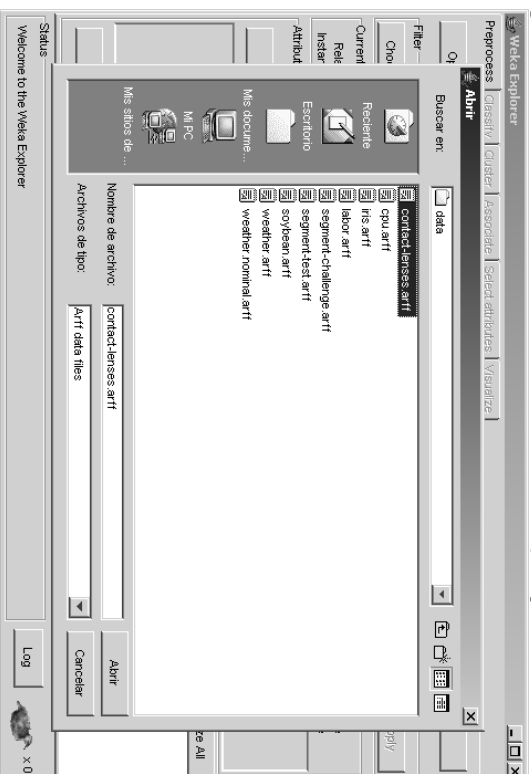
- Iniciando el explorador de WEKA



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Conceptos básicos

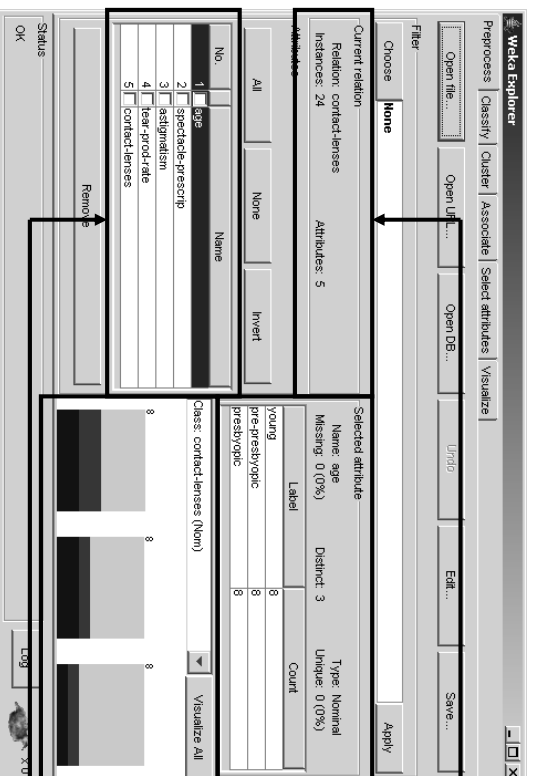
- Cargando datos en WEKA (Open file...)



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Conceptos básicos

- Observando los datos en WEKA



Datos generales de la colección

Datos del atributo seleccionado

Visualización del atributo seleccionado

Atributos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

El proceso de minería de datos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

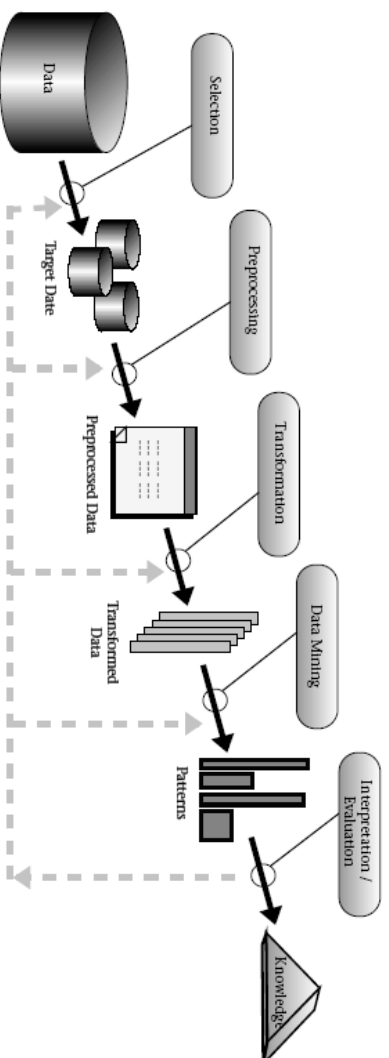
El proceso de minería de datos

- Más terminología
 - Descubrimiento de conocimiento en bases de datos .
Knowledge Discovery in Databases (KDD)
 - Desarrollo de técnicas y métodos para extraer conocimiento (= información útil) a partir de grandes volúmenes de datos
 - Proceso de convertir datos en bajo nivel (demasiados para ser comprendidos y asimilados) en otras formas
 - más compactas (informe corto)
 - más abstractas (una aproximación o modelo de cómo se generan los datos)
 - más útiles (un modelo predictivo para estimar casos futuros)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

El proceso de minería de datos

- El proceso del KDD
 - Todas las fases son importantes



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

El proceso de minería de datos

- Más terminología
 - Minería de datos – *Data Mining*
 - El paso del proceso del KDD que consiste en aplicar sobre los datos, algoritmos de análisis y descubrimiento que producen determinados patrones y modelos
 - Es la parte más cercana a “aprender” en sentido abstracto

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

El proceso de minería de datos

- **WEKA da soporte a muchas fases del proceso**
 - Selección de ejemplares y atributos
 - Preprocesado manual de la colección
 - Transformaciones vía filtros
 - Minería de datos (clasificación, agrupamiento, etc.)
 - Evaluación y visualización

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

Selección de atributos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Selección de atributos

- Algunos atributos
 - Pueden ser irrelevantes
 - Discriminar a jugadores de baloncesto y nadadores en función del color de los ojos
 - Pueden no aportar información o introducir ruido
 - E.g. Si sus valores aparecen de manera equiprobable en todas las clases
- Conviene usar sólo los atributos más informativos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Selección de atributos

- Existen métricas de calidad de los atributos
 - Miden la capacidad predictiva de un atributo en función de la relación entre sus valores y los de la clase
 - Estadística y teoría de la información
- Ejemplos
 - Ganancia de Información (*Information Gain*)
 - χ^2 (“chi” al cuadrado)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Selección de atributos

- Se usan seleccionando los atributos con mayor valor predictivo
 - Por encima de un valor en la medida (e.g. cero)
 - Los mejor situados en un ranking (e.g. el 1% superior)
- Se puede ganar efectividad
- Se gana eficiencia
 - Menos atributos => más rápido, menos memoria

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Selección de atributos

- Ganancia de información
 - Concepto de teoría de la información basado en la entropía
 - Usado también en aprendizaje de reglas y de árboles de decisión (entre otros)
 - Muy usada en contextos de clasificación de texto

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Selección de atributos

- **Ganancia de información**
 - Entropía = impureza de una colección de ejemplos
 - Sea una colección E de ejemplos, N clases (C_1, \dots, C_N), y sea $P_i = P(C_i)$
 - La entropía $H(E)$ se mide como

$$H(E) = \sum_{i=1}^N -P_i \cdot \log_2(P_i)$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Selección de atributos

- **Ganancia de información**
 - E.g. En la colección “Contact Lenses”
 - Hay tres clases = soft (5), hard (4), none (15)
 - La entropía es

$$\begin{aligned} H(CL) &= -(5/24) \cdot \log_2(5/24) \\ &\quad - (4/24) \cdot \log_2(4/24) \\ &\quad - (15/24) \cdot \log_2(15/24) \\ &= 0,45 + 0,43 + 0,42 = 1,33 \end{aligned}$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Selección de atributos

- Ganancia de información
 - Reducción esperada en la entropía al separar los ejemplos de acuerdo con un atributo
 - Sea la colección E, un atributo A con M valores V_1, \dots, V_M , y los conjuntos E_i de ejemplares con valor de A igual a V_i
 - La ganancia de información de A respecto E es

$$IG(E, A) = H(E) - \sum_{i=1}^M \frac{|E_i|}{|E|} \cdot H(E_i)$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

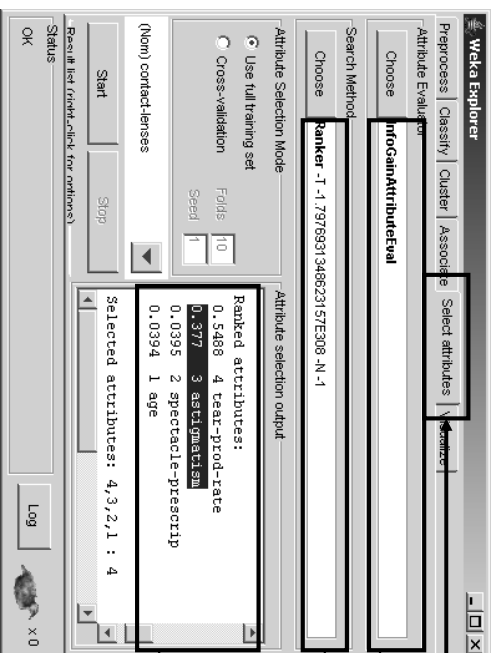
Selección de atributos

- Ganancia de información
 - E.g. En la colección “Contact Lenses”
 - El atributo *astigmatism* tiene 2 valores = yes (12), no (12)
 - La distribución de clases por valor es
 - yes = soft (0), hard (4), none (8)
 - no = soft (5), hard (0), none (17)
 - Las entropías de las sub-colecciones a = yes (E_1) y de a = no (E_2) son $H(E_1) = 0,92$ y $H(E_2) = 0,98$, luego
- $$IG(E, A) = H(E) - \left(\frac{|E_1|}{|E|}\right) \cdot H(E_1) - \left(\frac{|E_2|}{|E|}\right) \cdot H(E_2) = 1,33 - 0,5 \times 0,92 - 0,5 \times 0,98 = 0,37$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Selección de atributos

- Ganancia de información en WEKA



- Selección de atributos
- Métrica = Ganancia de Inf.
- Tipo de búsqueda = *ranker*
= producir un *ranking* de los atributos
- *Ranking* de atributos
astigmatism es el segundo mejor

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

PRISM: Inducción de reglas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

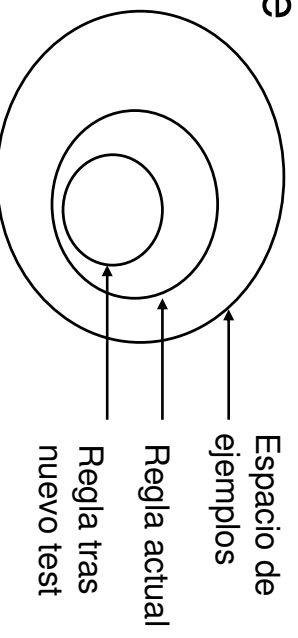
PRISM: Inducción de reglas

- Uno de los algoritmos más simples
- Algoritmo de recubrimiento (*covering*)
 - En cada paso, se construye una regla que cubre un subconjunto de ejemplares
 - Estrategia de “separa y vencerás” (*separate and conquer*)
 - Encuentra una regla útil, separa los ejemplos cubiertos, “vence” a los restantes
 - No “divide y vencerás”, porque los elementos cubiertos no se vuelven a examinar

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- Cada regla se construye agregando un test sobre un atributo
 - E.g. Age = young
- Los tests se seleccionan para maximizar la efectividad (porcentaje de acierto) de la regla
- Cada nuevo test reduce la cobertura



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- Selección del test
 - T = número de ejemplares cubiertos por la regla
 - P = número de ejemplos positivos (en la clase objetivo) cubiertos por la regla
 - Elegir el test que maximiza P/T
- Finalizar la regla cuando $P/T = 1$ o no se puede dividir más el conjunto de ejemplares

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- E.g. Empezamos la regla “IF test THEN hard”
 - Tests posibles
- | | P/T |
|---------------------------------------|------|
| Age = Young | 2/8 |
| Age = Pre-presbyopic | 1/8 |
| Age = Presbyopic | 1/8 |
| Spectacle prescription = Myope | 3/12 |
| Spectacle prescription = Hypermetrope | 1/12 |
| Astigmatism = no | 0/12 |
| Astigmatism = yes | 4/12 |
| Tear production rate = Reduced | 0/12 |
| Tear production rate = Normal | 4/12 |

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- E.g. Con el mejor test
IF Astigmatism = yes THEN hard
– Ejemplos cubiertos

Age	Spectacle prescription	Astigmatism	Tear production rate	Contact lenses
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- E.g. Continuamos con la regla
IF Astigmatism = yes AND *test* THEN hard
– Tests posibles
- | | P/T |
|---------------------------------------|-----|
| Age = Young | 2/4 |
| Age = Pre-presbyopic | 1/4 |
| Age = Presbyopic | 1/4 |
| Spectacle prescription = Myope | 3/6 |
| Spectacle prescription = Hypermetrope | 1/6 |
| Tear production rate = Reduced | 0/6 |
| Tear production rate = Normal | 4/6 |

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- E.g. En el siguiente refinamiento
IF Astigmatism = yes AND
Tear production rate = Normal THEN hard
– Ejemplos cubiertos

Age	Spectacle prescription	Astigmatism	Tear production rate	Contact lenses
Young	myope	yes	normal	hard
Young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	normal	none

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- E.g. Continuamos con la regla
IF Astigmatism = yes AND Tear production rate =
Normal AND *test* THEN hard
– Tests posibles P/T
Age = Young 2/2
Age = Pre-presbyopic 1/2
Age = Presbyopic 1/2
Spectacle prescription = Myope 3/3
Spectacle prescription = Hypermetrope 1/3
– En caso de empate => cobertura

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- E.g. En el siguiente refinamiento
IF Astigmatism = yes AND
Tear production rate = Normal AND
Spectacle prescription = Myope THEN hard
– Ejemplos cubiertos

Age	Spectacle prescription	Astigmatism	Tear production rate	Contact lenses
young	myope	yes	normal	hard
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
presbyopic	myope	yes	normal	hard

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- E.g. Regla final
IF Astigmatism = yes AND
Tear production rate = Normal AND
Spectacle prescription = Myope THEN hard
- Otra regla derivada sobre los ejemplos no cubiertos de la clase “hard”
IF Age = young AND
Astigmatism = yes AND
Tear production rate = normal THEN hard
- El proceso se repite para cada clase

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- **Pseudocódigo de PRISM**
 - Para cada clase C
 - Inicializar E al conjunto de ejemplares
 - Mientras E contiene ejemplares en la clase C
 - Crear regla nueva R con lado izdo vacío para clase C
 - Hasta que R es perfecta (o no quedan más atributos) hacer
 - Para cada atributo A no en R, y cada valor v,
 - Probar a agregar la condición $A = v$ al lado izdo de R
 - Seleccionar A y v to para maximizar P/T
 - (resolver empates con máximo P)
 - Agregar $A = v$ a R
 - Eliminar ejemplares cubiertos por R de E

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- **Las reglas de PRISM**
 - Se pueden aplicar sin orden explícito
 - Actúan como fragmentos de conocimiento independientes
- **Problemas**
 - Si son aplicables varias (con clases distintas)
 - Usualmente, elegir la clase más frecuente aplicable
 - Si no es aplicable ninguna
 - Usualmente, elegir la clase más frecuente (global)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

The screenshot shows the Weka Explorer interface with the PRISM classifier selected. The 'Classifiers' tree on the left shows 'PRISM' under the 'rules' category. The main window displays the following results:

Classified Instances	24	100	%
Classified Instances	0	0	%
DC	1		
error	0		
pred error	0	%	
rule error	0	%	
squared error	0	%	
of Instances	24		

Accuracy By Class ==

Rate	Precision	Recall	F-Measure	Class
1	1	1	1	soft
1	1	1	1	hard
1	1	1	1	none

Matrix ==

```
-- classified as
a = soft
b = hard
0 4 0 1
0 0 15 1 c = none
```

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

The screenshot shows the Weka Explorer interface with the WEKA classifier selected. The 'Classifiers' tree on the left shows 'WEKA' under the 'rules' category. The main window displays the following results:

Classified Instances	1102,47	100	%
Classified Instances	0	0	%
DC	1		
error	0		
pred error	0	%	
rule error	0	%	
squared error	0	%	
of Instances	1102,47		

Accuracy By Class ==

Rate	Precision	Recall	F-Measure	Class
1	1	1	1	soft
1	1	1	1	hard
1	1	1	1	none

Matrix ==

```
-- classified as
a = soft
b = hard
0 4 0 1
0 0 15 1 c = none
```

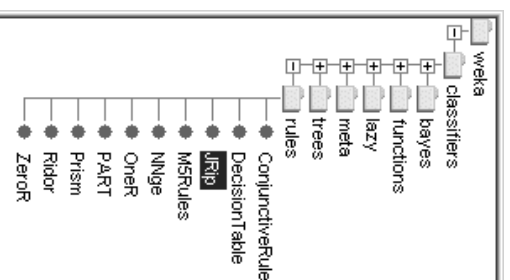
PRISM: Inducción de reglas

PRISM
en
WEKA

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

PRISM: Inducción de reglas

- Otros algoritmos de inducción de reglas
 - Ripper
 - Clásico de W. Cohen
 - Muy efectivo
 - PART
 - Reciente, usa árboles de decisión
 - Muy efectivo



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

Bayes Ingenuo

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- Sistema de aprendizaje basado en el teorema de Bayes
 - Modelado estadístico / probabilístico
 - Clasificador = tabla de probabilidades
 - Simple y efectivo
 - Aplica simplificaciones manifestamente falsas...
 - Pero los resultados son frecuentemente buenos
 - Bien fundamentado, estable
 - Pequeños cambios en los datos => pequeños cambios en el clasificador

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- Tres pasos
 - Pre-computar un conjunto/tabla de probabilidades
 - Averiguar la probabilidad de cada clase dado un ejemplar objetivo (sin clasificar)
 - Seleccionar la clase más probable

- E.g. Contact Lenses

$$e \in C \Leftrightarrow C = \operatorname{argmax}_{c \in \{\text{soft}, \text{hard}, \text{none}\}} (P(C = c | E = e))$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - Por el Teorema de Bayes
$$P(C = c|E = e) = \frac{P(E = e|C = c)P(C = c)}{P(E = e)}$$
 - Ergo, hay que computar
$$P(E = e|C = c) = \text{probabilidad del ejemplar } e \text{ dada } c$$
$$P(C = c) = \text{probabilidad de la clase } c$$
 - Pero no $P(E)$ – idéntico denominador para toda C

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - E.g. Cómputo de $P(C=\text{soft}|E=e)$ para $e = \langle \text{young, myope, yes, reduced} \rangle$
$$P(C = \text{soft}) = \frac{N(C = \text{soft})}{N}$$
 - Estimador de máxima verosimilitud = número de ejemplares (de entrenamiento) en *soft* dividido por el número total de ejemplares (de entrenamiento)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - E.g. Cómputo de $P(C=\text{soft}|E=e)$ para $e = \langle \text{young, myope, yes, reduced} \rangle$
$$P(E = e|C = \text{soft}) = P(\text{age} = \text{young}|C = \text{soft}) \times P(\text{spec} - \text{pre} = \text{myope}|C = \text{soft}) \times P(\text{astigs} = \text{yes}|C = \text{soft}) \times P(\text{tpr} = \text{reduced}|C = \text{soft})$$
 - Atributo = fragmento **INDEPENDIENTE** de evidencia = Bayes **INGENUO**

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - E.g. Cómputo de $P(C=\text{soft}|E=e)$ para $e = \langle \text{young, myope, yes, reduced} \rangle$
$$P(\text{age} = \text{young}|C = \text{soft}) = \frac{N(\text{age} = \text{young}|C = \text{soft})}{N(C = \text{soft})} = \frac{2}{5}$$
$$P(\text{tpr} = \text{reduced}|C = \text{soft}) = \frac{N(\text{tpr} = \text{reduced}|C = \text{soft})}{N(C = \text{soft})} = \frac{0}{5} = 0$$
 - **!!!!!!** Puede ocurrir $P(C=c|E=e)=0 \forall c!!!!!!$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - Para evitar $P(C=c|E=e)=0$, usar el estimador de Laplace (“sumar 1”)
 - Agregar 1 en el numerador, y el número de sucesos posibles en el denominador
 - E.g. age \in {young, pre-presbyopic, presbyopic} \Rightarrow tres sucesos
 - E.g. tpr \in {normal, reduced} \Rightarrow dos sucesos
 - A la larga converge al EMV = equi-probabilidad en caso de información nula

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - E.g. Cómputo de $P(C=\text{soft}|E=e)$ para $e = \langle \text{young, myope, yes, reduced} \rangle$

$$P(\text{age} = \text{young} | C = \text{soft}) = \frac{N(\text{age} = \text{young} | C = \text{soft}) + 1}{N(C = \text{soft}) + 3} = \frac{3}{8}$$

$$P(\text{tpr} = \text{reduced} | C = \text{soft}) = \frac{N(\text{tpr} = \text{reduced} | C = \text{soft}) + 1}{N(C = \text{soft}) + 2} = \frac{1}{7}$$

- Garantía $P(C=c|E=e) \neq 0 \forall c$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- En general, para calcular $P(C=c|E=e)$ para cualesquiera c, e , precisamos
 - $P(C=c) \forall c$
 - $P(C=c|A=a) \forall c, A, a$
- E.g. Contact Lenses
 - $P(C=soft), P(C=hard), P(C=none)$
 - $P(C=soft|age=young), P(C=soft|age=pre-presbyopic), P(C=soft|age=prebyopic), P(C=soft|spe-pre=myope), \dots$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses *

Clase (c)	P(C=c)	Clase (c)	P(C=c spe-pre=myope)	P(C=c spe-pre=hyper)
soft	0,22	soft	0,429	0,571
hard	0,18	hard	0,667	0,333
none	0,59	none	0,471	0,529

Clase (c)	P(C=c age=young)	P(C=c age=pre-presb)	P(C=c age=presb)
soft	0,375	0,375	0,250
hard	0,429	0,286	0,286
none	0,278	0,333	0,389

Clase (c)	P(C=c ast=no)	P(C=c ast=yes)	Clase (c)	P(C=c lpr=reduced)	P(C=c lpr=normal)
soft	0,857	0,143	soft	0,143	0,857
hard	0,167	0,833	hard	0,167	0,833
none	0,471	0,529	none	0,765	0,235

* Para $P(C=c)$ se usa el estimador de Laplace (sumar 1)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - E.g. Cómputo de $P(C=\text{soft}|E=e)$ para $e = \langle \text{young, myope, yes, reduced} \rangle$

$$\begin{aligned} P(C = \text{soft}|E = e) &= \frac{P(E = e|C = \text{soft})P(C = \text{soft})}{P(E = e)} \\ &= \frac{0,22 \times 0,37 \times 0,43 \times 0,14 \times 0,14}{0,000695} \\ &= \frac{P(E = e)}{P(E = e)} \end{aligned}$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - E.g. Cómputo de $P(C=c|E=e)$ para $e = \langle \text{young, myope, yes, reduced} \rangle$

$$P(C = \text{soft}|E = e) = \frac{0,000695}{P(E = e)}$$

$$P(C = \text{hard}|E = e) = \frac{0,007142}{P(E = e)}$$

$$P(C = \text{none}|E = e) = \frac{0,031223}{P(E = e)} \leftarrow \text{seleccionada}$$

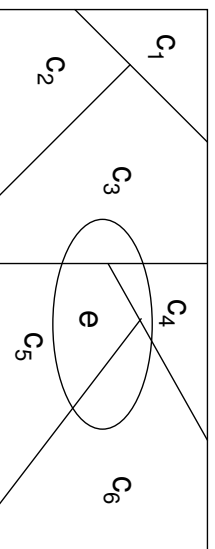
Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- Si precisamos auténticas probabilidades => Teorema de la Probabilidad Total
 - Si existen M clases c_1, \dots, c_M

$$P(E = e) = \sum_{i=1}^M P(E = e|C = c_i) \times P(C = c_i)$$

Espacio de sucesos
(clases, ejemplares)



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - E.g. Cómputo de $P(E=e)$ para $e = \langle \text{young, myope, yes, reduced} \rangle$
 - $P(E = e) = P(E = e|C = \text{soft}) \times P(C = \text{soft})$
 - $P(E = e|C = \text{hard}) \times P(C = \text{hard})$
 - $P(E = e|C = \text{none}) \times P(C = \text{none})$
 - Los términos coinciden con los numeradores anteriores => proyección al intervalo [0, 1]

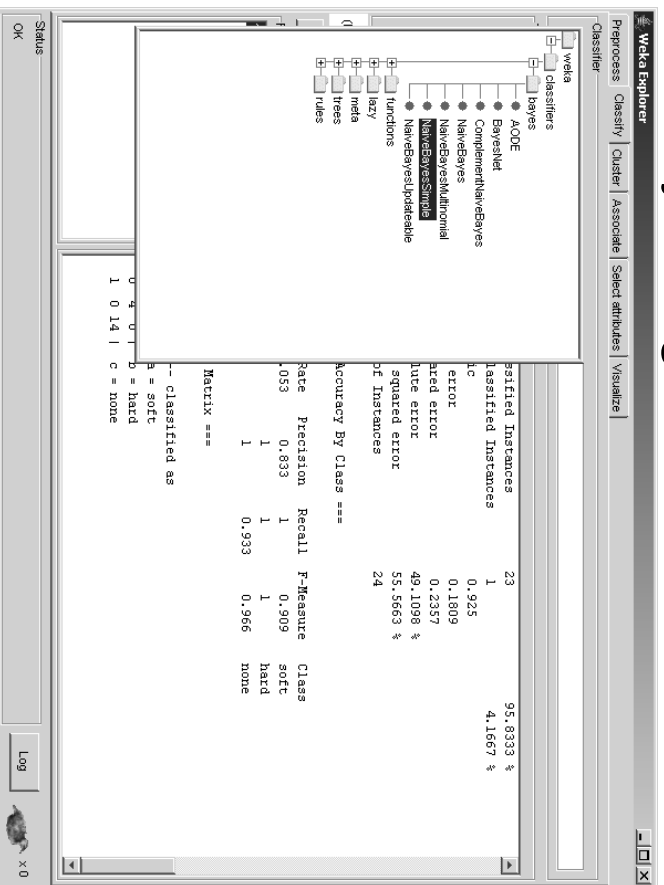
Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- E.g. Contact Lenses
 - E.g. Cómputo de $P(E=e)$ para $e = \langle \text{young, myope, yes, reduced} \rangle$
 $P(E = e) = 0,000695 + 0,007142 + 0,031223 = 0,03906$
 - $P(C = \text{soft} | E = e) = \frac{0,000695}{0,03906} = 0,0177$
 - $P(C = \text{hard} | E = e) = 0,1828$
 - $P(C = \text{none} | E = e) = 0,7993 \leftarrow \text{seleccionada}$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

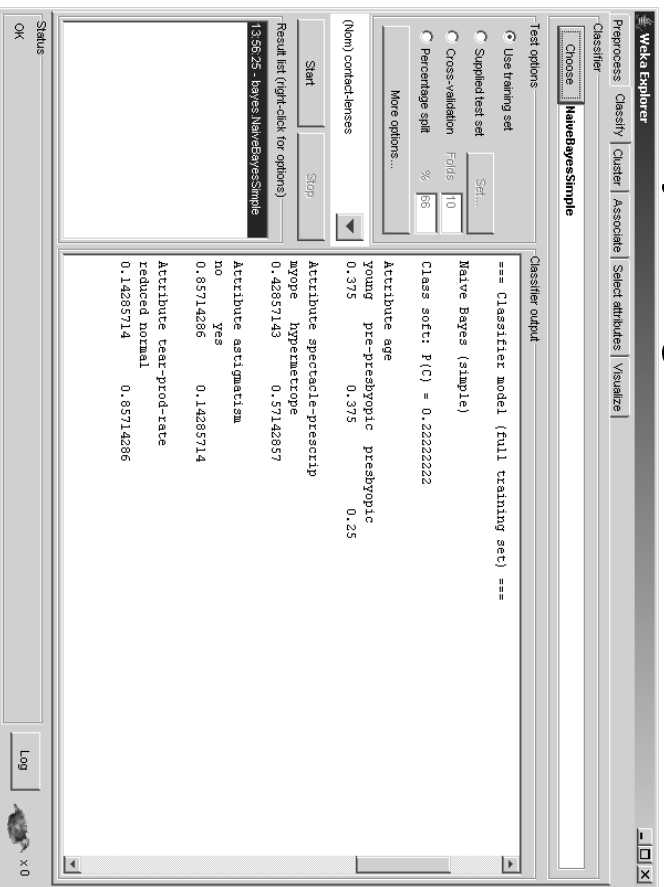


EN WEKA

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

En WEKA



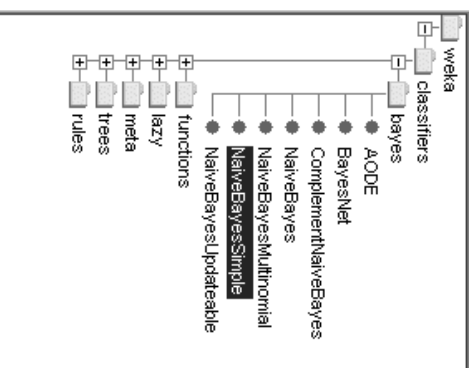
The screenshot shows the Weka Explorer interface with the NaiveBayesSimple classifier selected. The Classifier output pane displays the following results:

```
=== Classifier model (full training set) ===
Naive Bayes (simple)
Class soft: P(C) = 0.22222222
Attribute age
  young  pre-presbyopic  presbyopic
  0.375      0.375      0.25
Attribute spectacle-prescrip
  myope  hypermetrope
  0.42857143  0.57142857
Attribute astigmatism
  no      yes
  0.85714286  0.14285714
Attribute tear-prod-rate
  reduced normal
  0.14285714  0.85714286
```

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Bayes Ingenuo

- Versiones + avanzadas
 - NaiveBayes
 - Ampliación a atributos numéricos
 - Estimadores refinados
 - BayesNet
 - Redes de inferencia bayesiana
 - Reconocen explícitamente las dependencias entre atributos



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

ID3: árboles de decisión

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión

- Inducción de árboles de decisión = uno de los métodos más clásicos (Ross Quinlan)
- Árboles de decisión
 - Buena representación del conocimiento = operativa, clara y sencilla
- Algoritmo de inducción
 - Estrategia descendente y recursiva = divide y vencerás

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión

- Esquema del algoritmo
 - Seleccionar un atributo para el nodo raíz – crear una rama para cada posible valor del atributo
 - Separar los ejemplares en subconjuntos, uno por cada rama, según el valor del atributo
 - Repetir recursivamente para cada rama, usando el subconjunto asignado como colección
 - Detenerse si todas los ejemplares pertenecen a la misma clase o no hay más atributos
 - Clasificación = la clase más frecuente de la hoja

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión

- Elección del atributo de partición
 - Métrica de calidad del atributo => Ganancia de Información (IG)
 - Existen múltiples métricas con diferentes pero similares resultados en la efectividad
 - E.g. Ratio de ganancia (Gain Ratio) mejora los problemas producidos por atributos con muchos valores (ver + adelante)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

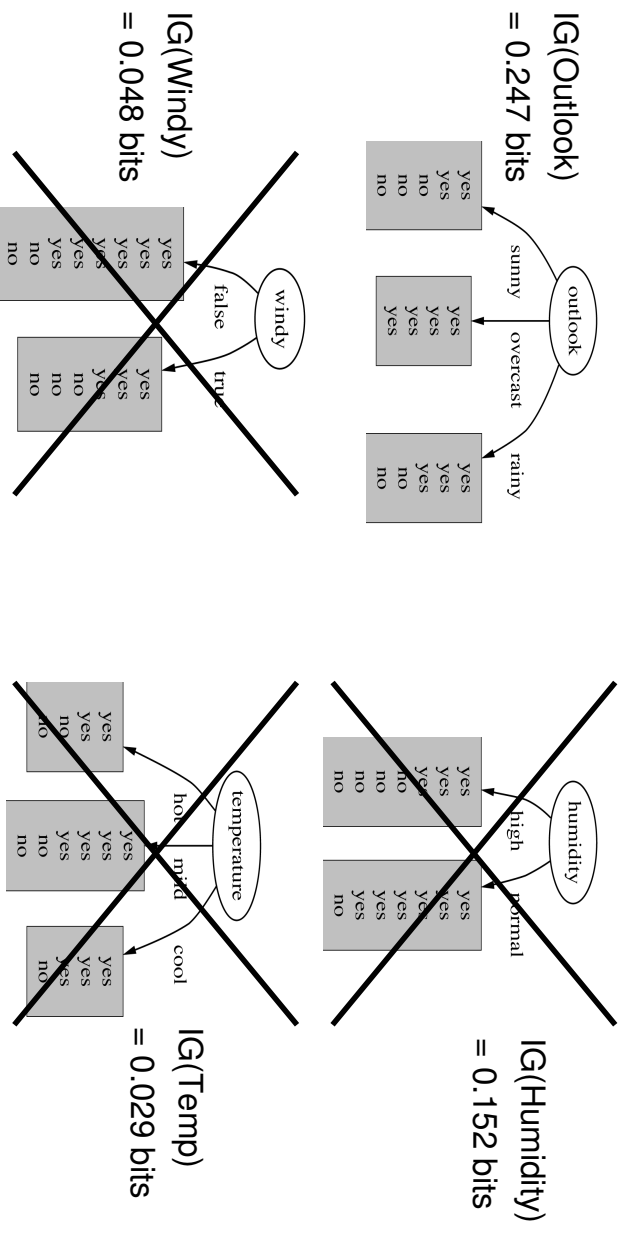
ID3: árboles de decisión

- E.g. Juego de tenis
 - Recomendar jugar o no al tenis según condiciones meteorológicas
 - 4 atributos y dos clases

Outlook	Temp	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

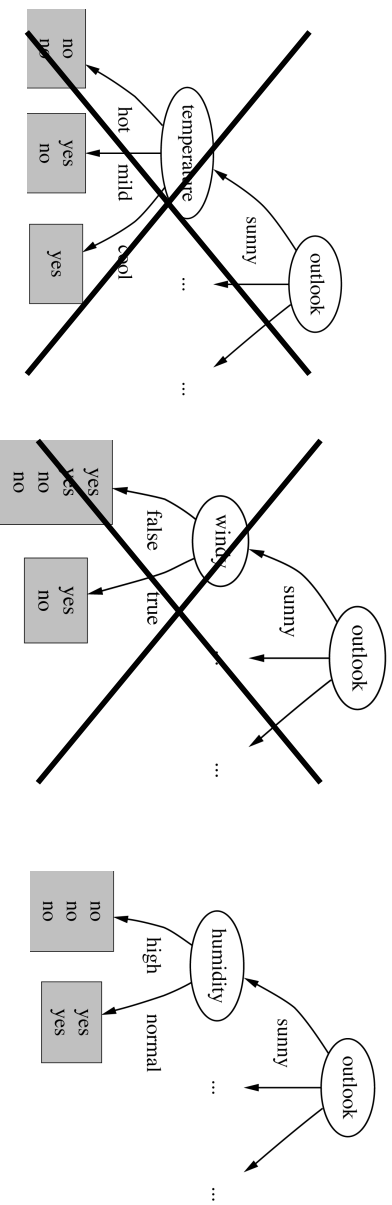
Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión



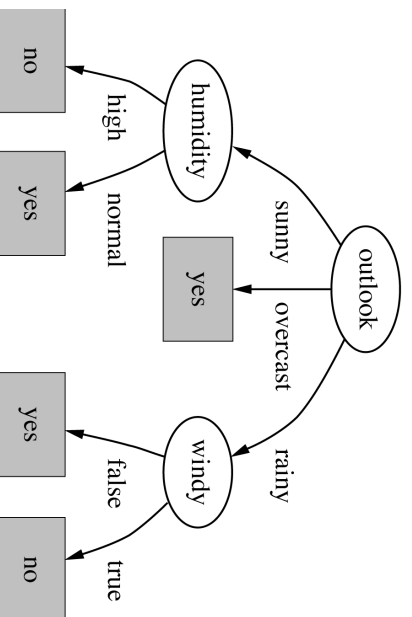
$IG(Temp) = 0.571$ bits

$IG(Windy) = 0.020$ bits

$IG(Humidity) = 0.971$ bits

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión



- No todas las hojas tienen que ser puras (sólo elementos de una clase)
 - El proceso se detiene cuando no es posible partir más

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión

- Los atributos con alto número de valores son problemáticos para IG
 - El caso extremo es un código de identificación único
 - IG muestra preferencia por ellos
 - Promueve el sobre-ajuste
 - Alta efectividad sobre los datos de entrenamiento...
 - Pero baja sobre los reales, operativos
 - Porque hemos particularizado demasiado

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión

- Otros algoritmos de inducción de árboles
 - C4.5 = ID3 mejorado con métricas de calidad distintas, tratamiento de atributos numéricos, poda del árbol para evitar el sobre-ajuste, etc.
 - Es J48 en WEKA
 - C5.0 = C4.5 con mejoras propietarias
 - CART similar a los anteriores
- En general, todos (los mejorados) son equivalentes en efectividad

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión

WEKA

The screenshot shows the WEKA Explorer interface. In the 'Classifiers' tree, 'ID3' is selected under 'DecisionStump'. The main window displays the following performance metrics:

Accuracy By Class ===			
Rate	Precision	Recall	F-Measure
1.2	0.889	0.889	0.889
1.111	0.8	0.8	0.8

Confusion Matrix ===			
	yes	no	classified as
yes	14	0	
no	0	1	

Summary			
Split	Folds	%	Mean absolute error
10	10	98	0.1429

Classifier output			
Test mode:	10-fold cross-validation		
=== Classifier model (full training set) ===			
ID3			
outlook = sunny			
humidity = high: no			
humidity = normal: yes			
outlook = overcast: yes			
outlook = rainy			
windy = TRUE: no			
windy = FALSE: yes			
Time taken to build model: 0.01 seconds			
=== Stratified cross-validation ===			
=== Summary ===			
Correctly Classified Instances	12	85.7143 %	
Incorrectly Classified Instances	2	14.2857 %	
Kappa statistic	0.6889		
Mean absolute error	0.1429		

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

ID3: árboles de decisión

WEKA

The screenshot shows the WEKA Explorer interface. In the 'Classifiers' tree, 'ID3' is selected under 'DecisionStump'. The main window displays the following performance metrics:

Accuracy By Class ===			
Rate	Precision	Recall	F-Measure
1.2	0.889	0.889	0.889
1.111	0.8	0.8	0.8

Confusion Matrix ===			
	yes	no	classified as
yes	14	0	
no	0	1	

Summary			
Split	Folds	%	Mean absolute error
10	10	98	0.1429

Classifier output			
Test mode:	10-fold cross-validation		
=== Classifier model (full training set) ===			
ID3			
outlook = sunny			
humidity = high: no			
humidity = normal: yes			
outlook = overcast: yes			
outlook = rainy			
windy = TRUE: no			
windy = FALSE: yes			
Time taken to build model: 0.01 seconds			
=== Stratified cross-validation ===			
=== Summary ===			
Correctly Classified Instances	12	85.7143 %	
Incorrectly Classified Instances	2	14.2857 %	
Kappa statistic	0.6889		
Mean absolute error	0.1429		

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Breve Introducción al Aprendizaje Automático con WEKA

Evaluación y visualización

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Evaluación de algoritmos

- Es importante evaluar la calidad del aprendizaje
 - Efectividad – Grado de acierto
 - Eficiencia
 - Tiempo invertido en aprender, en clasificar nuevos ejemplares, memoria
 - Claridad del conocimiento obtenido (clasificador)
 - Una regla es más sencilla de entender que una tabla de probabilidades, etc.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Evaluación de algoritmos

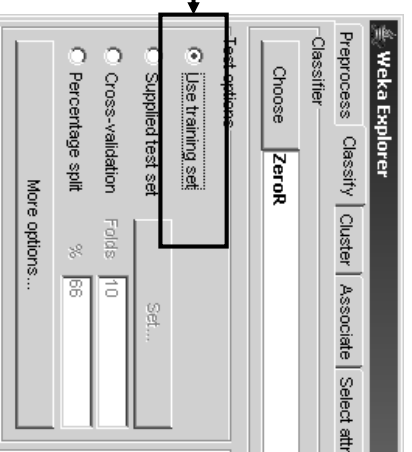
- Nos concentramos en efectividad
- La evaluación se compone de
 - Protocolo
 - Procedimiento de evaluación
 - ¿Cómo se tratan los datos?
 - Medidas o métricas
 - Sus valores definen la calidad del sistema
 - ¿Qué se calcula?

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Evaluación de algoritmos

- Protocolo 1 : colección de entrenamiento
 - Computar la medida objetivo sobre la propia *colección de entrenamiento*
 - El más simple
 - Injusto, no generalizable

Usar colección de entrenamiento →



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

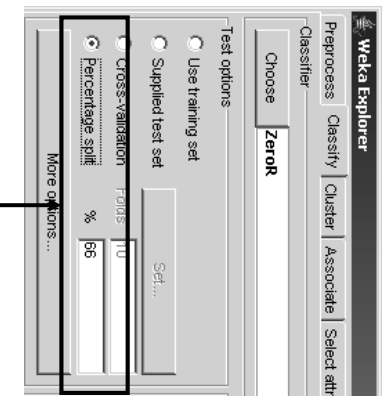
Evaluación de algoritmos

- Protocolo 2: (sub) colección de evaluación
 - Sobre una colección de evaluación separada
 - Se puede tomar del entrenamiento
 - Proporción – usualmente 66/33, 90/10
 - Se pierden datos de entrenamiento
 - Pero no para el entrenamiento final
 - Más justo, relativa generalidad

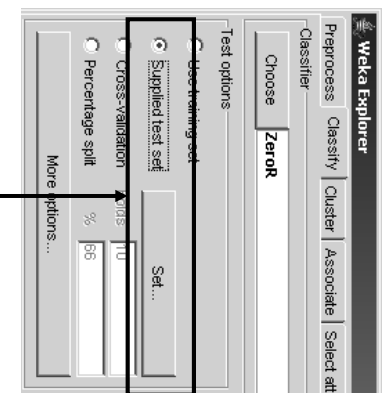
Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Evaluación de algoritmos

- Protocolo 2: (sub) colección de evaluación



Extraída de la colección
de entrenamiento



Disponible por otros
medios

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

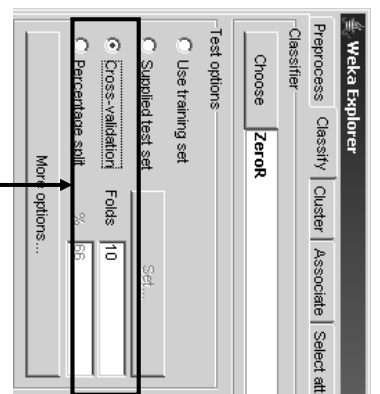
Evaluación de algoritmos

- Protocolo 3: Validación cruzada en K carpetas
 - Procedimiento
 - Se divide la colección de entrenamiento en K partes
 - Aleatoria, conservando la proporción entre clases
 - En K turnos, se reserva una parte para evaluar y se entrena sobre las K-1 restantes
 - Se promedian o acumulan los resultados de cada turno
 - El más justo y generalizador
 - Frecuentemente $K = 3, 5, 10$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Evaluación de algoritmos

Protocolo 3: Validación cruzada en K carpetas

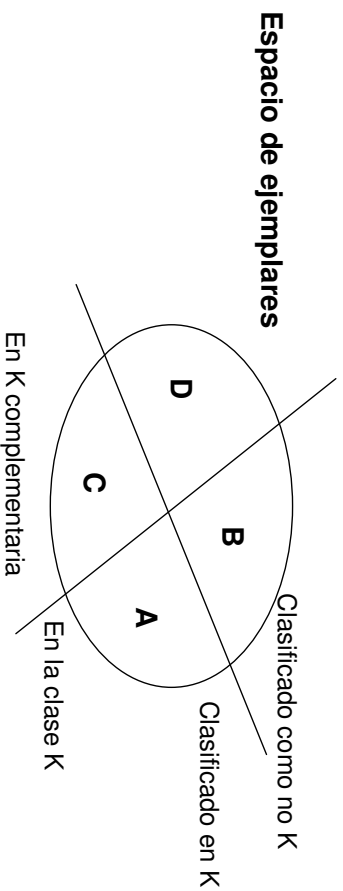


Opción y K (carpetas o grupos)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Evaluación de algoritmos

- Métricas de evaluación
 - Exactitud – *accuracy*
 - Porcentaje de aciertos sobre número de intentos
 - La más habitual



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Evaluación de algoritmos

- Tabla/matriz de contingencia/confusión

	Clasificado como K	Clasificado como no K
En K	A	B
No en K	C	D

$$accuracy = \frac{A + D}{A + B + C + D}$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

Evaluación de algoritmos

- Múltiples métricas
 - Aplicables en situaciones que lo requieran
 - Error, error cuadrático medio, cobertura, precisión, tasa de falsos positivos, etc.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid

The screenshot shows the Weka Explorer interface with the 'Classifier output' window open. The 'Prism' classifier has been used on the 'ruks-Prism' dataset. The output is as follows:

```
==== Summary ====
Correctly Classified Instances 13      54.1667 %
Incorrectly Classified Instances 7      29.1667 %
Kappa Statistic              0.3204
Mean absolute error          0.1944
Root mean squared error      0.441
Relative absolute error       63.0915 %
Root relative squared error   112.93 %
Unclassified Instances       24
Total Number of Instances    24

==== Detailed Accuracy By Class ====
TP Rate  FP Rate  Precision  Recall  F-Measure  Class
0.5      0.125     0.5        0.5     0.5        soft
0.333    0.118     0.333     0.333  0.333     hard
0.769    0.429     0.769     0.769  0.769     none

==== Confusion Matrix ====
 a b c <-- classified as
2 1 1 | a = soft
0 1 2 | b = hard
2 1 10| c = none
```

Accuracy

Error

Medidas para
cada clase

Tabla de confusión

Evaluación de algoritmos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – U. Europea Madrid