MEMORIA PRÁCTICAS 6-7

SISTEMAS DE LA INFORMACIÓN

Roldán Urueña, Diego Raúl (841723)

Romeo Lancina, Abel (846088)

Moreno Muñoz, Pablo (841972)

Escuela de
Ingeniería y Arquitectura
Universidad Zaragoza

Diego Raúl Roldán, Abel Romeo y Pablo Moreno
841723, 846088 y 841723

Índice

Introducción

En estas prácticas 6 y 7 de la asignatura de Sistemas de la Información de 3º de Ingeniería Informática se van a analizar distintos conjuntos de datos mediante Lucene y Weka.

Lucene es una biblioteca de software de código abierto desarrollada en Java, diseñada principalmente para facilitar la indexación y búsqueda eficiente de información en grandes conjuntos de datos de texto, mientra que Weka es un conjunto de herramientas de software de código abierto diseñado para realizar tareas de minería de datos y aprendizaje automático.

Diego Raúl Roldán, Abel Romeo y Pablo Moreno
841723, 846088 y 841723

Test libreria Lucene

Resultados

Buscando Contaminaci□n: Encontrados 1 hits.
1. ./ficheros/uno.txt 1.6400275

Buscando cambio clim□tico: Encontrados 3 hits.
1. ./ficheros/cuatro.txt 1.3217045
2. ./ficheros/tres.txt 1.0441608
3. ./ficheros/dos.txt 0.94759953

Buscando por: Encontrados 4 hits.
1. ./ficheros/uno.txt 0.20090158
2. ./ficheros/tres.txt 0.19296822
3. ./ficheros/cuatro.txt 0.19090943
4. ./ficheros/dos.txt 0.15896153

Buscando aeropuerto: Encontrados 1 hits.
1. ./ficheros/uno.txt 2.3586044

Preguntas

1. ¿Qué pasa si utilizamos el “StandardAnalyzer” en lugar del “SimpleAnalizer”?

Se omiten las palabras que se encuentran en el fichero stopwords.txt, se ha detectado que si se ejecuta con esta condición en la consulta de la palabra “por” no se encuentran resultados en ningún fichero

2. ¿Qué función tiene el fichero “stopwords.txt”?

Como se ha comentado anteriormente, contiene las palabras que se omitiran si se ejecuta con la opción StandardAnalyzer

3. ¿Qué ocurre si en la búsqueda ponemos “contaminacion” o “cambio climatico” (sin tildes)?

En el caso de contaminacion sin tilde no encuentra ninguna coincidencia ya que la opción de Simple Analyzer convierte los caracteres en minúscula y sabe diferenciar entre caracteres con tilde y sin ella, por ello no encuentra ningún contaminacion sin tilde
En el caso de cambio climatico, encuentra solo los resultados que da la palabra cambio por separado ya que es la única que encuentra.

4. ¿Y si hacemos esta búsqueda utilizando el “SpanishAnalizer”? ¿Por qué ocurre esto?

En este caso encuentra los resultados tanto con tilde como sin tilde.

5. ¿Qué ocurre si re-indexamos todos los ficheros cada vez que ejecutamos el programa, en lugar de, simplemente, reabrir el índice creado previamente?

Si los reindexamos se añadirán los ficheros del directorio al índice, en este caso la función crearIndiceEnUnDirectorio() recorre la colección ficherosAIndexar y se añade cada fichero al índice sin realizar una verificación previa de cambios en los ficheros.
Por lo tanto se realizará la búsqueda varias veces sobre el mismo fichero.

Clasificador de flores IRIS

El clasificador ZeroR es un clasificador muy simple que se utiliza comúnmente como línea de base para comparar con otros clasificadores más complejos. La característica principal del clasificador ZeroR es que predice la clase más frecuente en el conjunto de datos para todas las instancias. En otras palabras, ignora por completo las características o atributos del conjunto de datos y simplemente predice la clase mayoritaria.

En este caso, el clasificador ZeroR está prediciendo solo una clase para todas las instancias y tiene un 33% de precisión en un conjunto de datos con tres clases. Esto significa que una de esas clases es la clase mayoritaria y representa aproximadamente el 33% de las instancias. El clasificador ZeroR es útil como referencia inicial, pero su simplicidad a menudo resulta en una baja precisión.

El clasificador J48 construye un árbol de decisión a partir del conjunto de datos de entrenamiento. Cada nodo interno del árbol representa una pregunta sobre un atributo, y cada hoja representa la predicción de la clase. Durante el entrenamiento, el árbol se construye de manera que las preguntas en los nodos maximizan la separación entre las clases.

El clasificador J48 ha obtenido un rendimiento del 96%, lo cual significa que ha logrado construir un árbol de decisión que se adapta bien al conjunto de datos y es capaz de realizar predicciones precisas. Cabe destacar que, aunque el rendimiento es excelente, es posible que el modelo esté sobreajustando los datos de entrenamiento, reduciendo la precisión al interpretar nuevos datos.

Reglas de asociación

Al realizar el análisis de asociación de atributos, obtenemos una serie de reglas que relacionan distintos valores del dominio de uno o más atributos con otro atributo distinto. Aquí un ejemplo:

```
humidity=normal windy=FALSE ==> play=yes
```

En este caso, vemos que la regla indica que cuando la humedad es normal y no es un día ventoso, se puede jugar.
Se muestran las reglas de mejores a peores, es decir, las que son apoyadas por más instancias dadas en el dataset, por lo tanto las que tenemos que tener más en cuenta son estas primeras.
Este dataset analiza los días que si se ha podido jugar a tenis teniendo en cuenta una serie de atributos, por lo tanto podríamos decir que solo nos interesan las reglas que tienen como resultado (en la parte derecha). Esto no es del todo correcto ya que si nuestro atributo resultado es x y las reglas son las siguientes:

```
1  y ==> x
2  z ==> y
```

Si consideramos correctamente la segunda regla podríamos llegar a otra regla resultante de las anteriores:

```
3  z ==> x
```

Por lo tanto es importante tener en cuenta todas las reglas posibles para que no se nos puedan escapar relaciones de este tipo.

En el caso del dataset de los partidos de tenis podríamos sacar las siguientes conclusiones acerca de las posibilidades de jugar un día sabiendo el estado del tiempo general (soleado, nubes o lluvia), la temperatura, la humedad y si es día ventoso o no. Según el análisis asociativo, si es un día nublado, se podrá jugar sin

problemas, de la misma forma, si es un día con humedad normal y sin viento también se podrá jugar. Si el día es lluvioso y no hay viento se podrá jugar, pero en cambio si el día está soleado y la humedad ambiental es alta, no se podrá jugar.

Resultados

i. Precisión al detectar las opiniones positivas.	0.824
ii. Precisión al detectar las opiniones negativas.	0.794
iii. Recall al detectar las opiniones positivas.	0.784
iv. Recall al detectar las opiniones negativas.	0.832
v. Precisión promedio.	0.809
vi. Recall promedio.	0.808

Se han procesado los datos eliminando del vector de palabras que contengan caracteres que no sean los del abecedario con la expresión regular de Perl `\w*[^A-Za-z\s]\w*`, para eliminar palabras irrelevantes para el algoritmo.
Los resultados del clasificador no mejoran apenas, la precisión promedio aumenta en 2 milésimas y el recall aumenta en 3 milésimas.

v. Precisión promedio.	0.811
vi. Recall promedio.	0.811

Profundización en los árboles de clasificación

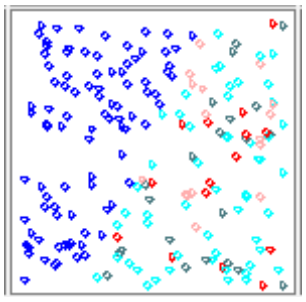
Al igual que al clasificar las flores IRIS, en este caso siempre se escogerá la clase mayoritaria, es decir, “drugY” con el algoritmo ZeroR. Pero en este caso acierta el 45.5% de las veces ya que esta distribución no es equitativa.
Si en cambio se escoge el algoritmo J48, la precisión superará el 90%, al igual que en las flores IRIS.

```
K <= 0.055221
|   K <= 0.037124: drugY (56.0)
|   K > 0.037124
|   |   Na <= 0.685143
|   |   |   BP = HIGH
|   |   |   |   Na <= 0.656371: drugA (6.0)
|   |   |   |   Na > 0.656371: drugY (2.0/1.0)
|   |   |   BP = LOW
|   |   |   |   Sex = F: drugC (3.0)
|   |   |   |   Sex = M: drugX (4.0/1.0)
|   |   |   BP = NORMAL: drugX (11.0/1.0)
|   |   Na > 0.685143: drugY (33.0/2.0)
K > 0.055221
|   BP = HIGH
|   |   Age <= 50: drugA (17.0)
|   |   Age > 50: drugB (15.0)
|   BP = LOW
|   |   Cholesterol = HIGH: drugC (14.0/1.0)
|   |   Cholesterol = NORMAL: drugX (13.0)
|   BP = NORMAL: drugX (26.0)
```

Este es el árbol de decisión seguido para J48 utilizando el conjunto de datos "training set".

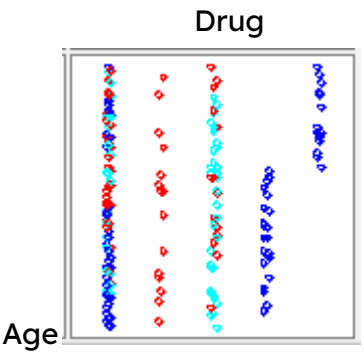
Si se observan las gráficas de relación entre pares de variables definido para un campo, por ejemplo, la siguiente gráfica relaciona Sodio (Na) - Potasio (K), por lo que para unos valores por encima de la diagonal principal, se suministrará casi siempre "drugY".

K



Na

Otro ejemplo es el siguiente, que relaciona Age (Edad) - Drug para la presión sanguínea. Se observa que "drugB", que es la de la última columna, es suministrada a los pacientes de edad avanzada y con alta presión sanguínea (puntos azules), en cambio, a aquellos cuyas edad va desde joven a mediana y con alta presión sanguínea, les es suministrado "drugA".



Retoque de variables

Tras convertir los datos K y Na en uno solo debido a su visible dependencia, el árbol de decisión se ha simplificado, como se puede observar en la imagen inferior, así como ha mejorado notablemente la precisión, siendo esta de un 100% al utilizar J48 con el "training set", y de un 99% al emplear una validación cruzada con 10 grupos.

```
Na_to_K <= 0.0666: drugY (91.0)
Na_to_K > 0.0666
|   BP = HIGH
|   |   Age <= 50: drugA (23.0)
|   |   Age > 50: drugB (16.0)
|   BP = LOW
|   |   Cholesterol = HIGH: drugC (16.0)
|   |   Cholesterol = NORMAL: drugX (18.0)
|   BP = NORMAL: drugX (36.0)
```

Cronograma

Gestión de horas

Actividad	Roldán Urueña, Diego	Romeo Lancina, Abel	Moreno Muñoz, Pablo	Total
Pruebas Lucene	1 hora	0.5 horas	3.5 horas	**5 horas**
Pruebas Weka	2 horas	2.5 horas	0.5 horas	**5 horas**
Redacción memoria	2 horas	3 horas	2 horas	**7 horas**
Total	**5 horas**	**6 horas**	**6 horas**	**17 horas**

Dificultades encontradas

En las pruebas de la librería Lucene la dificultad radica en comprender a fondo el comportamiento de diferentes analizadores, como el "StandardAnalyzer", el "SimpleAnalyzer" y el "SpanishAnalyzer", y en anticipar cómo estos afectarán la indexación y la recuperación de información. Además, la gestión de stopwords y la sensibilidad a las tildes en el idioma español agrega una capa adicional de complejidad.

El apartado de Weka está bien explicado en el guion la práctica, no ha habido dificultades, salvo por la profundización de los árboles de clasificación, donde la única dificultad encontrada ha sido el entender las gráficas para poder explicarlas.

Conclusión

En la fase de pruebas de la librería Lucene, se llevaron a cabo búsquedas utilizando distintos analizadores y se obtuvieron resultados significativos. El análisis de las búsquedas proporcionó valiosa información sobre el comportamiento de los analizadores y su impacto en los resultados.
Para las pruebas con Weka se han empleado diferentes clasificadores para analizar conjuntos de datos que permitan identificar patrones para predecir eventos dependientes a estos.

En resumen, se ha aprendido a analizar datos mediante las herramientas proporcionadas y gracias a la precisión del guion se ha amenizado la carga de trabajo, pues están perfectamente detallados los pasos a seguir para la realización de esta práctica.