INTELIGENCIA ARTIFICIAL EN LAS ORGANIZACIONES

PRÁCTICA 2

(Data Mining)

Grado en Ingeniería Informática

Campus de Colmenarejo

Curso 2019/2020

Autores

Paula Gilabert Robles - 100363657@alumnos.uc3m.es

Alvaro Gonzalez Muñoz - 100363552@alumnos.uc3m.es

Eduardo Ureña Toledano - 100329937@alumnos.uc3m.es

Índice

Introducción

En esta práctica intentaremos clasificar un conjunto de reseñas de hoteles en 5 categorías (*Poor*, *Fair*, *Good*, *Very Good* y *Excellent*).

En primer lugar, desarrollaremos un poco el contexto de la práctica.

En segundo lugar, en la "**Parte I**", explicaremos cómo hemos realizado el preprocesado de los datos necesario para realizar la segunda parte.

En tercer lugar, en la "**Parte II**", escogeremos uno de los conjuntos de datos preprocesados y lo analizaremos con la función de clustering, utilizando diferentes valores, y posteriormente elegiremos uno de estos modelos de clustering y le aplicaremos diferentes algoritmos de generación de árboles y reglas para poder analizar cuál es el algoritmo que mejores resultados obtiene con el cluster escogido.

Contexto de la práctica

La práctica se basa en la técnica "Text Mining" que forma parte del "Data Mining".

El "Data Mining" consiste en encontrar patrones que expliquen el comportamiento de un conjunto de datos en un contexto específico, siendo dicho contexto en nuestro caso los datos extraídos de las reseñas de hoteles de una gran cantidad de huéspedes.

El "Text Mining" es la parte del "Data Mining" que implica el procesamiento de texto de documentos, que es lo que tenemos que hacer en esta práctica.

El proceso común para hacer minería de datos suele constar de los siguientes pasos:

Determinar los objetivos: Se especifican los objetivos que se van a tener en cuenta a la hora de analizar un determinado conjunto de datos.

Preprocesar los datos: Esta etapa suele consumir alrededor del 70% del tiempo total de un proyecto de "Data Mining", y consiste en preparar los datos para que puedan ser analizados de la manera más satisfactoria posible.

Elegir el modelo más adecuado: Para saber qué modelo se debería elegir, se empieza haciendo un análisis estadístico de los datos. Teniendo en cuenta este análisis y los objetivos planteados, se elegirán uno o varios algoritmos pertenecientes al campo de la Inteligencia Artificial.

Analizar los resultados: Se estudian los resultados obtenidos y se comparan con los estadísticos en vista de encontrar una cierta coherencia y decidir si son lo suficientemente buenos para poder tomar decisiones en base a ellos.

Parte I: Preprocesado

Lo primero que hemos tenido que hacer es usar una Macro en Excel para generar un archivo de texto con cada reseña y su número de estrellas, generando 500 archivos por cada una de las 5 estrellas posibles, teniendo así un total de 2500 reseñas (luego se asignará una categoría a cada reseña en función de dicho número de estrellas).

Después hemos tenido que introducir un comando en la interfaz SimpleCLI de Weka, para recopilar en un fichero .arff las 2500 reseñas con su categoría correspondiente.

Finalmente hemos tenido que usar el filtro "*StringToWordVector*" para convertir el archivo en una serie de palabras donde cada una es un atributo, eliminando las stop words con el archivo "*StopWordsEN.txt*" y las derivaciones de las palabras, en la medida de lo posible, con el stemmer "Iterated Lovins", que es el que ha dado mejores resultados.

Parte II: Clustering

En la siguiente tabla podemos ver el reparto que hay de atributos entre los diferentes clusters en porcentajes. Las casillas amarillas son las que se corresponden con los mejores valores.

	Euclides	**Manhattan**
2 clusters seed 10	0 382 (15%) 1 2118 (85%)	0 1 (0%) 1 2499 (100%)
5 clusters seed 10	0 1 (0%) 1 895 (36%) 2 535 (21%) 3 795 (32%) 4 274 (11%)	0 1 (0%) 1 652 (26%) 2 364 (15%) 3 1274 (51%) 4 209 (8%)
10 clusters seed 10	0 1 (0%) 1 161 (6%) 2 33 (1%) 3 484 (19%) 4 74 (3%) 5 529 (21%) 6 1 (0%) 7 566 (23%) 8 454 (18%) 9 197 (8%)	0 1 (0%) 1 164 (7%) 2 3 (0%) 3 557 (22%) 4 65 (3%) 5 480 (19%) 6 1 (0%) 7 390 (16%) 8 723 (29%) 9 116 (5%)
2 clusters seed 50	0 1 (0%) 1 2499 (100%)	0 1 (0%) 1 2499 (100%)

5 clusters seed 50	0 1 (0%) 1 382 (15%) 2 1 (0%) 3 1 (0%) 4 2115 (85%)	0 1 (0%) 1 121 (5%) 2 1 (0%) 3 1 (0%) 4 2376 (95%)
10 clusters seed 50	0 1 (0%) 1 6 (0%) 2 1 (0%) 3 1 (0%) 4 852 (34%) 5 697 (28%) 6 2 (0%) 7 139 (6%) 8 800 (32%) 9 1 (0%)	0 1 (0%) 1 1 (0%) 2 1 (0%) 3 1 (0%) 4 1242 (50%) 5 257 (10%) 6 1 (0%) 7 4 (0%) 8 991 (40%) 9 1 (0%)

En la siguiente tabla podemos observar la suma de los errores en el modelo de Euclides y la suma de las distancias en el modelo de Manhattan. (Las casillas amarillas son las que se corresponden con los mejores valores).

SUMA DE ERRORES/DISTANCIAS	**Euclides**	**Manhattan**
2 clusters seed 10	37817.420280123646	78154.84240355906
5 clusters seed 10	36722.26011402287	76508.18191442752
10 clusters seed 10	35740.1849321207	75390.96836901983
2 clusters seed 50	38959.43785136125	78111.4750996825
5 clusters seed 50	37753.6061144346	77463.7721071518
10 clusters seed 50	36571.54060771499	76486.82033968923

Como podemos observar, en función de los clusters obtenemos que algunos modelos tienen la mayoría de clusters vacíos, mientras que otros tienen los datos más repartidos entre ellos, por lo que serán mejores para su análisis mediante los algoritmos de árboles.

Los modelos que mejores datos han otorgado en cuanto a la repartición de atributos han sido, tanto de 5 como de 10 clústeres, y de la seed 10.

En cuanto a la suma de errores, los resultados son bastante elevados, pero si consideramos los menores de todos los modelos, coincide que los modelos de 10 clústeres y seed 10 son los que dan valores más bajos, al igual que en las agrupaciones son los que están más repartidos.

Teniendo en cuenta lo anterior, hemos decidido seleccionar como modelos a analizar los modelos que tienen 10 clusters y seed 10.

Manhattan	**zeroR**	**OneR**	**PART**	**Hoeffding Tree**	**J48**	**RandomTree**
Nº correctos	1651.0	1651.0	1651.0	1651.0	1651.0	1651.0
Nº incorrectos	849.0	849.0	849.0	849.0	849.0	849.0
Nº sin clasificar	245.0	585.0	500.0	585.0	562.0	272.0
correctos (%)	604.0	264.0	349.0	264.0	287.0	577.0
incorrectos (%)	0.0	0.0	0.0	0.0	0.0	0.0
sin clasificar (%)	28.857479387514722	68.90459363957598	58.89281507656066	68.90459363957598	66.19552414605418	32.037691401649
Media error absoluto	0.0	0.0	0.0	0.0	0.0	0.0
Error cuadrático medio	0.0	0.5963927884580755	0.4833082806413024	0.5963927884580755	0.5731859839408081	0.14115829926295784
Error absoluto relativo	0.15982708700748777	0.0621908127208483	0.08675148437229935	0.09420501055597903	0.07905424943588953	0.135983510011177986
Error cuadrático relativo	0.2826555699159916	0.24938085876997115	0.2722904335597515	0.21382064646836207	0.23981927304329537	0.368579708038653153

Euclides	zeroR	OneR	PART	HoeffdingTree	J48	RandomTree
Nº correctos	1649.0	1649.0	1649.0	1649.0	1649.0	1649.0
Nº incorrectos	851.0	851.0	851.0	851.0	851.0	851.0
Nº sin clasificar	192.0	642.0	606.0	642.0	606.0	302.0
correctos (%)	659.0	209.0	245.0	209.0	245.0	549.0
incorrectos (%)	0.0	0.0	0.0	0.0	0.0	0.0
sin clasificar (%)	22.561692126909517	75.4406580493537	71.2103407755581 7	75.4406580493537	71.21034077555817	35.487661574618095
Media error absoluto	0.0	0.0	0.0	0.0	0.0	0.0
Error cuadrático medio	0.0	0.6901907708172211	0.6507685788034057	0.6901907708172211	0.6490644861574921	0.20481367886794635
Error absoluto relativo	0.1645357126920143	0.04911868390129266	0.0583660597309 9204	0.07599610679780364	0.059547963836182514	0.12902467685076502
Error cuadrático relativo	0.286800464900082	0.22162735368472156	0.2309307139037 7717	0.19090408281692714	0.2245774788593262	0.35846563542295995

Como podemos observar en ambas tablas, el algoritmo que da mejores resultados es el *zeroR*. Sin embargo, el que da el mejor valor en el "Error absoluto relativo" es el *RandomTree* y el del "Error cuadrático relativo" es el *HoeffdingTree*.

A pesar de que en ambos modelos los mejores resultados coinciden en cuanto al algoritmo que los proporciona, varían en los valores numéricos, por lo que el mejor modelo es el de *Euclides*, ya que en comparación con el de *Manhattan*, casi todos sus valores son mejores excepto los de "Nº correctos" y "Nº incorrectos".