Práctica 2
Análisis de las reseñas de Tripadvisor

Inteligencia Artificial en las Organizaciones

Grupo 83-1

Miguel Gutiérrez Pérez
100383537@alumnos.uc3m.es

Mario Lozano Cortés
100383511@alumnos.uc3m.es

Alba Reinders Sánchez
100383444@alumnos.uc3m.es

Alejandro Valverde Mahou
100383383@alumnos.uc3m.es

GitHub: ***InteligenciaArtificialOrganizaciones***

31 de octubre de 2020

Índice

1. Introducción

La siguiente sección incluye

2. Parte 1: Clasificación

2.1. Análisis y preprocesado de datos

Explicación de los pasos previos de lo que vimos en clase y división en cat

2.2. Experimentación

Generalidades de todos los experimentos. hacer hincapié en lo de las stopwords. Caso base

2.2.1. Experimentación básica

Tablas de cada uno de lo que hicimos esa tarde

2.2.2. Experimentación avanzada

Combinación de los mejores resultados básicos

2.3. Comentario de los resultados obtenidos

Mucho text

3. Parte 2: Clustering

En la segunda parte de esta práctica se va a realizar una aproximación mediante ***clustering***, pero antes de esto, recordar que esta técnica consiste en agrupar instancias sin etiquetar de manera que las instancias pertenecientes aun mismo grupo sean más similares entre sí que con las de otro grupo diferente.

En este caso, se agruparán las instancias procesadas que obtuvieron un mejor resultado en la primera parte de la práctica, esta agrupación se hará con el algoritmo ***K-Medias***, que se basa en el valor medio de las distancia de cada grupo para generar los grupos.

Este proceso se vuelve a realizar en *Weka*, y se compone de los siguientes pasos:

- Cargar el archivo *.arff* con los datos generado en la parte anterior.
- Generar diferentes modelos a partir de estos datos con *K-Medias* y compararlos.
- Analizar el mejor modelo obtenido.
- Ejecutar diversos algoritmos de generación de reglas y árboles de decisión con el mejor modelo obtenido.

3.1. Experimentación

Los experimentos que se llevan a cabo son los que se muestran en la en la Figura X. Los parámetros del algoritmo que se modifican son la *seed* (10, 20 y 30), el *números de clusters* (2, 3, 4, 5 y 6) y el *tipo de distancia* (Euclidea y Manhattan).

ID Experimento	Error	Seed	Número de clusters	Tipo de distancia
0	25436,31171	10	5	Euclidea
1	25382,55604	20	5	Euclidea
2	25218,09711	30	5	Euclidea
3	25703,79957	10	2	Euclidea
4	25704,73815	20	2	Euclidea
5	25704,73815	30	2	Euclidea
6	25545,30221	10	3	Euclidea
7	25684,13735	20	3	Euclidea
8	25643,51841	30	3	Euclidea
9	56512,35917	10	5	Manhattan
10	56666,77425	20	5	Manhattan
11	56680,27896	30	5	Manhattan
12	25509,57983	10	4	Euclidea
13	25670,38789	20	4	Euclidea
14	25449,76246	30	4	Euclidea
15	25250,01674	10	6	Euclidea
16	25342,88323	20	6	Euclidea
17	25160,2531	30	6	Euclidea

Figura 1: Tabla de experimentos con *K-Medias*

Como se puede apreciar en la Figura X, se prueba para un mismo número de clusters las 3 seeds. Cabe destacar que los experimentos probados con distancia de *Manhattan* cometen mucho más error que los que usan la *Euclidea*. Por este motivo en la gráfica donde se comparan los errores no se añaden, ya que no son relevantes.

Los errores cometidos por los experimentos que usan distancia *Euclidea* varían en el rango de 25160.25, siendo este el mejor, a 25704.73, siendo este el peor. De estos, los mejores resultados son los marcados en verde.

Este error representa la suma de las distancias medias a cada cluster, por ello cuanto menor sea el error menor serán estas distancias y como consecuencia las instancias estarán mejor agrupadas.

A continuación se muestra en la Figura X la gráfica comparativa de los errores para poder visualizar mejor los errores de cada experimento.
Por lo tanto, se concluye con que el mejor resultado es el obtenido en el **Experimento 17**, así que en la siguiente sección se va a analizar en detalle este modelo.

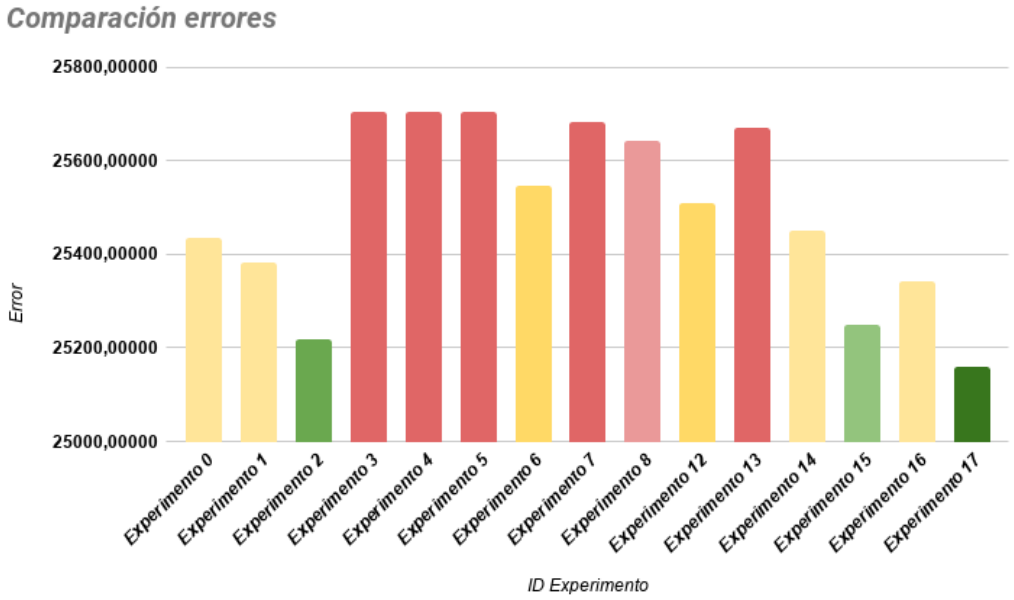

Figura 2: Gráfica comparativa de los errores

3.2. Mejor modelo

El mejor modelo es el obtenido con el Experimento 17, se compone de **6 clusters** aunque según la gráfica de la Figura X, parece que deja 2 clusters vacíos. Mientras que el cluster 4 es el que tiene mayor número de elementos seguido del cluster 5. Por otro lado los clusters 2 y 3 tienen muchos menos elementos.

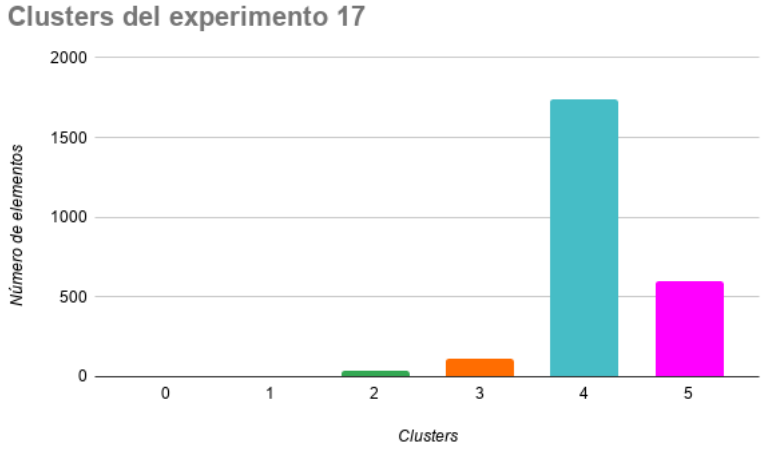

Figura 3: Gráfica clusters Experimento 17

Ya que con la Figura X no se puede afirmar que los clusters 0 y 1 estén vacíos, se van a visualizar los datos en la siguiente tabla de la Figura X.

Cluster	Número de elementos	Porcentaje de elementos
0	1	0
1	6	0
2	39	2
3	116	5
4	1737	69
5	601	24

Figura 4: Tabla clusters Experimento 17

Una vez que se tienen los números concretos se observa que los cluster 0 y 1 no están vacíos pero poseen muy pocos elementos, 1 y 6 respectivamente. Por lo tanto, se llega a la conclusión de que estos elementos son *outliers* y no aportan nada.

Tras este análisis del mejor modelo, se ejecutan los siguientes algoritmos de generación de reglas y árboles de decisión con este modelo para descibrir los clusters:

- *PART*
- *J48*
- *RandomForest*
- *JRip*

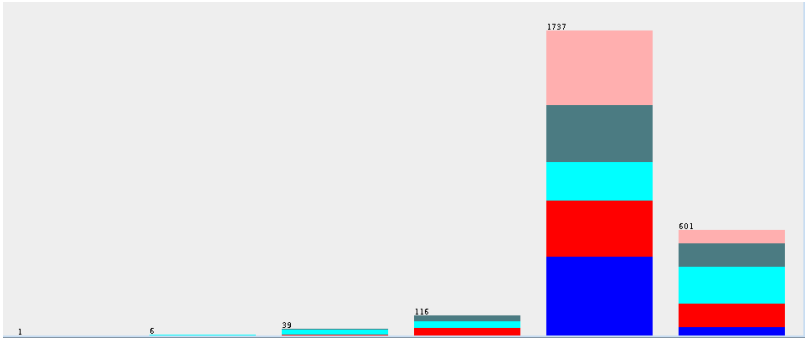

Figura 5: Gráfica clusters-clase Experimento 17

4. Conclusiones

5. Contexto de la práctica

6. Referencias

1. Introduction to Neurons in Neural Networks. Medium. Consultado en Octubre 2020. Url: https://medium.com/artificial-neural-networks

7. Anexos

1. Perceptron Multicapa usando '*K Fold*'
   perceptron_kfold.py
2. Perceptron Multicapa usando '*split percentage*'
   perceptron_split.py
3. Programa para realizar la predicción de los modelos
   predict.py
4. Tabla de resultados de los experimentos de la primera parte
   valores_reales_vs_predicciones_&_errores_absolutos_parte1.xlsx
5. Tabla de resultados de los experimentos de la segunda parte
   valores_reales_vs_predicciones_&_errores_absolutos_parte2.xlsx