SIW PL11-12. Datamining.

Hugo Fonseca Díaz
uo258318@uniovi.es

Escuela de Ingeniería Informática. Universidad de Oviedo.

10 de diciembre de 2020

1. Objetivo

El objetivo de esta práctica es crear un clasificador que realice predicciones de las bajadas o subidas del índice bursátil *Dow Jones* según una serie de noticas internacionales registradas ese día. Para crear dicho clasificador se usará la herramienta *Orange* [1].

2. Diseño

Para dicho clasificador se ha creado un corpus formado por los datos extraídos de una hoja de cálculo. Al preprocesarse dicho corpus se le añade una lista de palabras vacías incluidas en el fichero *stopwords.txt* manualmente creado. Después se forma una bolsa de palabras con una frecuencia de términos y un cálculo del IDF. Dicha bolsa se conecta con varios modelos que a su vez muestran los resultados de sus cálculos en un namograma. Este namograma muestra las palabras más importantes a la hora de clasificar si una noticia hará que el índice *Dow Jones* baje o suba. Además, la bolsa de palabras y el modelo con el que se ha decidido clasificar, el *bosque aleatorio*, se conectan a un nodo que prueba y evalua el clasificador.

3. Resultados

Una vez observados los namogramas obtenidos, se prueban y clasifican los diferentes modelos. La regresión logística muestra un porcentaje de acierto por debajo del 30 % (sobre el 28 %). El bayesiano ingenuo obtiene un porcentaje de acierto de casi el 50 % (sobre un 49 %). Al probar con otros modelos se ha llegado a la conclusión de que el mejor de ellos es el *bosque aleatorio*, con un porcentaje de acierto de algo mas del 50 %. Al no poder visualizarlo con un namograma no se puede observar que términos considera más importantes, por lo que simplemente se observa el resultado de las pruebas mediante una matriz de confusión.

4. Conclusión

Al analizar los resultados experimentales se observa una tasa de acierto demasiado baja, un 50 % no es un porcentaje lo suficientemente grande como para considerar utilizar dicho clasificador para una tarea como la presente. Este porcentaje tan bajo puede explicarse por varios factores. El primero es que quizás los datos otorgados no son lo suficientemente significativos como para resolver el problema, por lo que quizás habría que buscar un mejor dataset. El segundo es debido al propio funcionamiento del índice bursátil, ya que al ser tan complejo no se puede asegurar que haya una relación de causalidad entre el tipo de noticias y la subida o bajada del mismo.

Referencias

[1] Página de *Orange*, `https://orange.biolab.si/`. Última vez accedido 10 de diciembre de 2020.