Universidad Carlos III de Madrid

Aprendizaje Automático

Grado en Ingeniería Informática

Grupo 83

Práctica 2: Aprendizaje basado en instancias

Autores:
Daniel Medina García
Alejandro Rodríguez Salamanca

2 de abril de 2016

Índice

Introducción

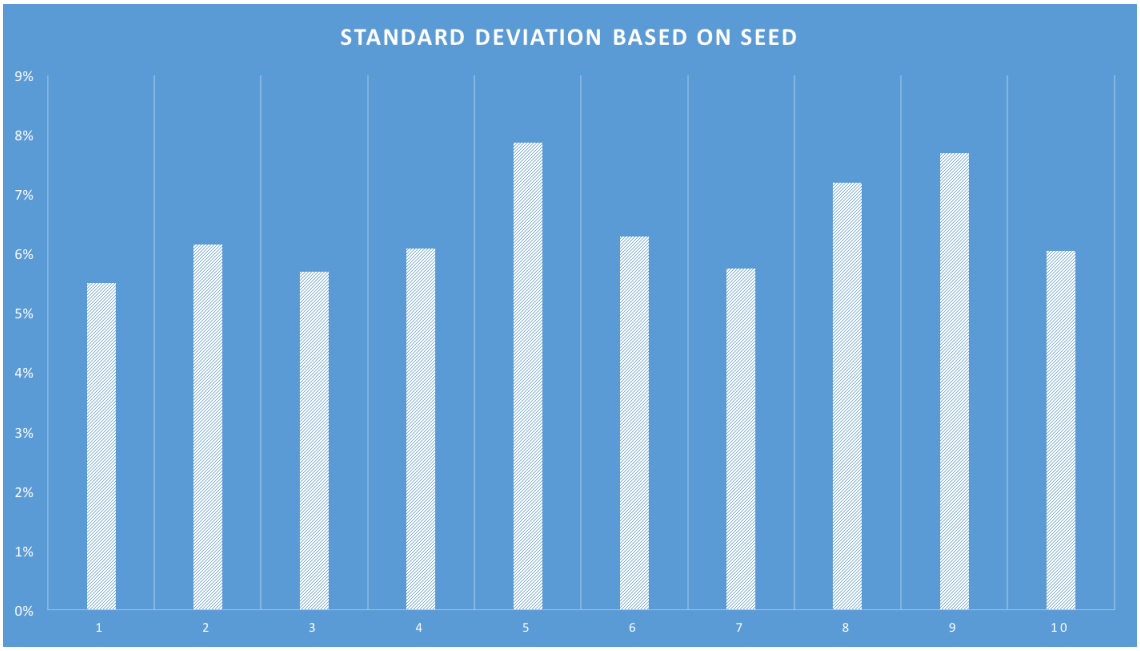
El presente documento contiene la memoria del trabajo realizado para esta segunda práctica de Aprendizaje Automático. En esta práctica el equipo ha utilizado el aprendizaje basado en instancias, haciendo uso de la técnica de *clustering* para poder implementar funciones de afinidad y agilizar así la clasificación.

1. Recogida de información

2. Clustering

Tras probar todos los diferentes algoritmos de clustering ofrecidos por Weka, hicimos un primer filtro con aquellos que nos daban un número manejable de clusters (o se podía configurar dicho número) para evitar aquellos que generaban demasiados (menos de un 5 % de pertenencia) o insuficientes (menos de 5). Esta primera selección nos dejó con Cobweb, EM, FarthestFirst y SimpleKMeans. Comparando los algoritmos, buscamos dos propiedades: equilibrio entre los clusters y "estabilidad.^entre ejecuciones con modificamiento en los parámetros (i.e. semilla u otras constantes). Esta comparativa nos hizo decantarnos por SimpleKMeans y EM, pues los porcentajes de pertenencia a cada cluster eran más parecidos entre sí y distintas semillas resultaban en clusters de dimensiones similares.

Si bien los resultados eran parecidos entre estos dos algoritmos, el elevado coste en tiempo para elaborar el clustering con EM nos hizo decantarnos por SimpleKMeans. Mostramos a continuación la justificación de nuestra decisión, donde observamos el equilibrio conseguido con este algorimo de clustering y su estabilidad ante el cambio de la semilla. Cabe destacar que con los otros algoritmos encontramos variaciones muy superiores (e.g. 11 y 17 % con FarthestFirst), alejadas de la media de 6 % obtenida con SimpleKMeans.



Para potenciar la eficacia de la clusterización, probamos a normalizar los datos. Sin embargo, los resultados obtenidos fueron los mismos. Sin embargo, la normalización de los datos nos ayudará en la función de pertenencia a clasificar más fácilmente por lo que decidimos mantenerla en nuestro modelo.

3. Generación del agente automático

3.1. ¿Por qué ha sido útil realizar clustering previa de las instancias?

Ahorra bastante tiempo en comparaciones para la clasificación. Al tener clusters ya hechos, sólo compararemos la instancia nueva con aquellas que pertenezcan al mismo cluster en lugar de con todo el set de entrenamiento.

3.2. ¿Por qué es importante usar pocos atributos en técnicas de aprendizaje no supervisado?

3.3. ¿Qué ventaja tiene el uso del aprendizaje basado en instancias con respecto al visto en la práctica 1?

3.4. ¿Consideras que el agente funcionaría mejor si se introdujesen más ejemplos? ¿Por qué?

Un conjunto de entrenamiento más grande podría ayudar a formar clusters más informados. Sin embargo, también haría más numerosas las instancias en cada cluster, ralentizando así el proceso de clasificación posterior. Este drowback podría contrarrestarse añadiendo un mayor número de clusters.

4. Evaluación de los agentes

5. Conclusiones

Problemas encontrados

Comentarios personales