

Inteligencia artificial: Práctica 8

Álvaro García Tenorio *

Miguel Pascual Domínguez **

8 de mayo de 2018

Índice general

Índice general	1
1 Agrupamiento	2
1.1. Descripción del conjunto	2
1.2. Parametrización del algoritmo de agrupamiento	3
1.3. Resultados	3
1.4. Conclusiones y respuestas	3
2 Clasificación	4
2.1. Descripción del conjunto	4
2.2. Parametrización del J48	5
2.3. Resultados	5
2.4. Conclusiones y respuestas	5
3 Regresión	8
3.1. Descripción del conjunto	8
3.2. Parametrización del perceptrón multicapa	8
3.3. Parametrización del K-NN	8
3.4. Resultados	8
3.5. Conclusiones	8

* alvgar14@ucm.es

** miguepas@ucm.es

Capítulo 1

Agrupamiento

1.1. Descripción del conjunto

Para este experimento hemos seleccionado un conjunto de datos llamado ausencias de trabajo. Los datos pertenecen a un estudio académico realizado por la universidad Nove de Julho de Brasil, por dos estudiantes y un profesor, durante tres años desde 2007 hasta 2010 en una empresa brasileña. Los datos, los hemos obtenido de la siguiente pagina, <http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>.

Explicuemos ahora los atributos/variables de este conjunto de datos:

1. ID (Identificador): variable formada por un conjunto de números del 1 al 36 que consisten en 36 identificadores de 36 personas distintas.
2. Motivo de la ausencia: catalogado en 28 números que cada uno corresponde con un motivo de incidencia, los 21 primeros se tratan de las primeras 21 enfermedades del código internacional de enfermedades (ICD) y los 7 ultimos, corresponden con 7 motivos no incluidos en el ICD, que son el acompañamiento de otro paciente (22), una consulta médica (23), donación de sangre (24), análisis médico (25), ausencia injustificada (26), fisioterapia (27) y dentista (28).
3. Mes de ausencia: número real que corresponde con el mes de la ausencia del 1 al 12.
4. Día laboral de la semana: lunes (2), martes (3), miércoles (4), jueves (5) y viernes (6).
5. Estación del año: verano (1), otoño (2), invierno (3) y primavera (4).
6. Gasto de transporte: número real.
7. Distancia desde casa hasta la oficina (km): número real.
8. Tiempo de trabajo: número entero.
9. Edad: número entero
10. Media de carga de trabajo: número real.
11. Porcentaje de cumplimiento de trabajos: número real entre 0.0 y 100.0.
12. Incidencia disciplinaria: Si (1) y No (0).

13. Educación: número real, catalogado de la siguiente manera; bachillerato (1), graduado (2), posgraduado (3) y master y doctorado (4).
14. Número de hijos: número real.
15. Bebedor social: Si (1) y No (0).
16. Fumador social: Si (1) y No (0).
17. Número de mascotas: número real.
18. Peso: número real.
19. Altura: número real.
20. Índice de masa corporal: número real.
21. Tiempo de ausencia en horas: número real.

1.2. Parametrización del algoritmo de agrupamiento

Veamos ahora como preparamos los datos y las variables para aplicar el algoritmo de agrupamiento jerárquico.

Empecemos primero observando si tiene sentido o no estandarizar o normalizar o no hacerlas nada a las variables. Dado que algunas de ellas a pesar de que son números, realmente se les deberían considerar variables categóricas, por ejemplo el ID, el motivo de la ausencia, el día laboral, la estación del año, la incidencia disciplinaria, y el bebedor y fumador social; y por tanto al tener que considerarlas así, no tendría ningún sentido realizarles ajuste numérico.

Fijemonos ahora en el resto de las variables, AQUI FALTA DECIR PORQUE LAS OTRAS VARIABLES NO LAS ESTANDARIZAMOS NI NORMALIZAMOS PERO NO SE ARGUMENTAR PORQUE (PORQUE LA RAZON QUE SE ME OCURRE ES PURA "PEREZA").

Decidamos ahora el número de clusters que queremos para nuestro agrupamiento, para ello tenemos que observar que dependiendo de que variable tomamos como principal, se necesitará un número u otro, por ejemplo si cogemos como variable, el día laboral, lo interesante sería tomar como número de clusters igual a 5, para observar como el algoritmo te agrupa por cada día laboral o por ejemplo por el motivo de la ausencia, o la estación del año, etc. Por todo esto, hemos decidido optar por

1.3. Resultados

1.4. Conclusiones y respuestas

Capítulo 2

Clasificación

2.1. Descripción del conjunto

Para este tipo de problema, lo interesante es coger un conjunto/database que tenga pocos atributos y muchas instancias, por ello el conjunto que hemos seleccionado es un dataset sobre una evaluación de coches, hemos obtenido los datos de la siguiente página <http://archive.ics.uci.edu/ml/datasets/Car+Evaluation>. Los archivos descargados no venían en formato .arff, por tanto hemos tenido que añadir parte del código de definición de atributos y de la relación, pero los datos no han sido modificados.

Explicaremos ahora los atributos/variables, todos ellos categóricos, ninguno numérico; de este conjunto de datos:

1. Compra: Precio de compra; muy alto (vhigh), alto (high), medio (med) y bajo (low).
2. Mantenimiento: Precio de mantenimiento; muy alto (vhigh), alto (high), medio (med) y bajo (low).
3. Puertas: Número de puertas clasificado en 4 categorías; 2, 3, 4 y 5 o más.
4. Personas: Número de personas clasificado en 3 categorías; 2, 4 y más.
5. Maletero: Capacidad del maletero; pequeña, media y grande.
6. Seguridad: baja, media y alta.
7. Tipo de coche según aceptabilidad: inaceptable (unacc), aceptable (acc), bueno (good) y muy bueno (vgood).

Una vez comentado los atributos, expliquemos el problema.

Tal y como son los datos, hemos decidido que para aprovechar el potencial de los mismos, lo más lógico es plantear el problema de clasificar la aceptabilidad de los coches en función de los otros 6 atributos, ya que así podemos observar más adelante el poder de algunos atributos como la seguridad y el número de personas que ya clasifican una cantidad de los coches.

Mostramos aquí el histograma de la variable tipo de coche según aceptabilidad.

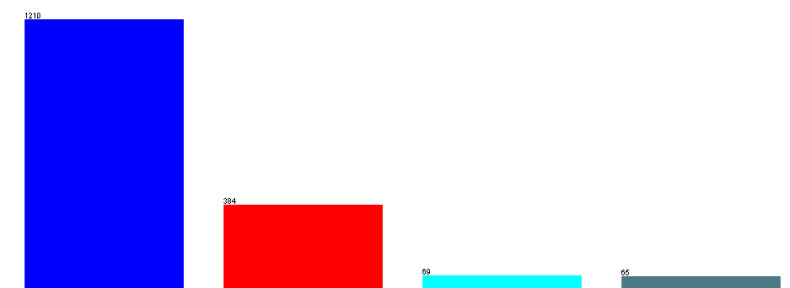


Figura 2.1: Histograma de la variable tipo de coche según aceptabilidad

2.2. Parametrización del J48

Dado que todas nuestras variables son categóricas no se les puede normalizar (englobar los valores entre 0 y 1) ni estandarizar (modificar los datos para que tengan media 0 y varianza 1). Primero se realizará una ejecución del algoritmo sin validación ni entrenamiento, después se realizará otra ejecución en validación y entrenamiento al 66 %.

En la primera ejecución se han mantenido los parámetros del algoritmo predeterminado por Weka y en la segunda ejecución al realizar un entrenamiento y una validación al 66 %, vimos que el árbol resultante era muy parecido al original sin entrenamiento ni validación, por tanto decidimos añadir una nueva peculiaridad a esta segunda ejecución y era que de manera forzada siempre se realizaran particiones binarias en el árbol, es decir que formara un árbol binario.

2.3. Resultados

Mostremos primero los resultados de la primera ejecución, primero mostramos el resumen de weka en la Figura 2.2, en el que está incluido los valores del recall, precisión los errores y la matriz de confusión; y después el árbol en la Figura 2.3.

Mostraremos ahora los resultados de la segunda ejecución, primero mostramos el resumen de weka en la Figura 2.4, en el que está incluido los valores del recall, precisión los errores y la matriz de confusión; y después el árbol en la Figura 2.5.

Comentemos ahora los resultados obtenidos en ambas ejecuciones:

2.4. Conclusiones y respuestas

Capítulo 3

Regresión

- 3.1. Descripción del conjunto
- 3.2. Parametrización del perceptrón multicapa
- 3.3. Parametrización del K-NN
- 3.4. Resultados
- 3.5. Conclusiones