Árboles de Decisión

García Prado, Sergio

2 de marzo de 2017

Resumen

En este documento se analiza el comportamiento de los algoritmos de generación de árboles de decisión ID3 *y* J48 *desde el punto de vista de la discretización de atributos (en el caso de* ID3 *de forma manual). Para ello se ha utilizado el conjunto de datos* Thoracic Surgery Data Data Set*[3]. La herramienta utilizada para el aprendizaje automático ha sido* WEKA*[2].*

1. ¿Por qué no se puede aplicar directamente ID3?

El conjunto de datos utilizado[3] está formado por **470** instancias, las cuales describen resultados acerca de la esperanza de vida tras operaciones de cáncer pulmonar. El conjunto de datos presenta atributos con las siguientes características:

- 10 atributos de tipo categórico binario.
- 1 atributo de tipo categórico con 3 valores distintos.
- 1 atributo de tipo categórico con 4 valores distintos.
- 1 atributo de tipo categórico con 7 valores distintos.
- 1 atributo de tipo numérico real en el rango [1,44, 6,3].
- 1 atributo de tipo numérico real en el rango [0,96, 86,3].
- 1 atributo de tipo numérico entero en el rango [21, 87].
- 1 atributo de tipo categórico binario, que representa el valor de la clase.

La razón por la cual no se puede aplicar el algoritmo *ID3* sobre este conjunto de datos es la existencia de atributos de naturaleza numérica, para los cuales dicho algoritmo no está diseñado.

2. Realice las modificaciones previas en el fichero de datos para que pueda llevar a cabo lo anterior y proporcione los resultados aplicando el método de retención o Hold-Out para la formación del experimento

Para poder aplicar el algoritmo *ID3* existen distintas alternativas, como la discretización de los atributos numéricos o la eliminación de los mismos. Nótese que la eliminación no es una buena alternativa debido a que de esta manera se deja de utilizar información valiosa para la tarea de clasificación, por tanto, en este caso se ha decidido realizar una discretización.

El método que se ha utilizado para dicha tarea ha sido el más básico posible, realizando una partición de 10 intervalos sobre el rango de valores que toma cada uno de los atributos numéricos.

3. Volviendo al fichero original, pase el algoritmo J48 también aplicando el método de retención o Hold-Out. Analice el árbol obtenido sin poda. ¿Se podría prescindir de algún atributo? Si es así, hágalo y compare los resultados de nuevo

4. Habrá notado que cuando se usa algún atributo numérico, implícitamente se aplica una discretización al plantear las diferentes ramas del árbol a partir de él. ¿Por qué es más eficiente esta técnica que la aplicada en 2?

5. Plantee, entonces, una discretización basada en el punto anterior, aunque no resulte ser binaria, sino en tantos tramos como induzcan los valores usados al formar las ramas del árbol con J48 sin poda (Hold-Out). Introduzca este fichero de nuevo al algoritmo ID3. Compare los resultados con el J48 de 3

Referencias

[1] Calonge Cano, T., and Alonso GonzáLez, C. J. Técnicas de Aprendizaje Autómatico, 2016/17.

[2] The University of Waikato. Weka. `http://www.cs.waikato.ac.nz/ml/weka/`.

[3] UCI Machine Learning Repository. Thoracic Surgery Data Data Set. `http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data`.