Árboles de Decisión

García Prado, Sergio

4 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*[4]. La herramienta utilizada para el aprendizaje automático ha sido* WEKA*[3]. La metodología experimental que se ha seguido a lo largo del documento ha sido un* Hold-Out *de* $\frac{1}{3}$ *para tareas de prueba.*

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

El conjunto de datos utilizado[4] está formado por **470** instancias, las cuales describen resultados acerca de la esperanza de vida tras operaciones de cáncer pulmonar. Dicho conjunto 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 el mismo es la existencia de atributos de naturaleza numérica, para los cuales dicho algoritmo no está diseñado. Por lo tanto, para que la clasificación con este algoritmo pueda llevarse a cabo, será necesaria una tarea de adaptación previa del conjunto de datos a los requisito de *ID3*

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, entre ellas se encuentran la discretización de los atributos numéricos o la eliminación de los mismos. Nótese que la eliminación no es la mejor alternativa en la mayoría de casos, 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 en **4 intervalos de la misma anchura** sobre el rango de valores que toma cada uno de los atributos numéricos. Tras procesar el conjunto de datos con el algoritmo *ID3* teniendo en cuenta la metodología experimental citada anteriormente, el árbol de decisión generado se ilustra en la figura 4. Tal y como se puede apreciar, este árbol presenta muchas ramificaciones. La causa de ello es que por cada atributo numérico, se han generado 4 ramas y dado que *ID3* genera todas las posibles combinaciones posibles a partir de las instancias destinadas a la fase de entrenamiento, el fenómeno combinatorio produce dicho efecto.

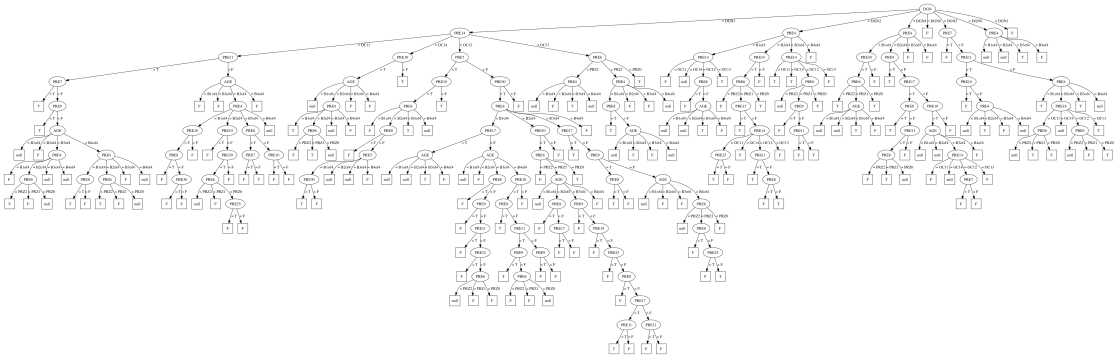

Figura 1: *Árbol de decisión generado a partir del algoritmo* ID3

La matriz de confusión resultante de dicha clasificación se muestra en la tabla 1. Nótese que se han utilizado **160 instancias** pero el árbol de decisión ha dejado sin clasificar 5 de ellas, por lo que la tasa de acierto global es del 72,5 %.

		Valor Real		
		Positivo	Negativo	p_j
Valor Predicho	Positivo	5	18	0,2173
	Negativo	21	111	0,8409
	π_j	0,1677	0,8322	$N = 155$

Tabla 1: *Matriz de confusión del conjunto de datos discretizado previamente y después entrenado por el algoritmo* ID3

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

Puesto que el algoritmo *J48* si que permite el procesado de conjunto de datos con atributos continuos, en este caso no es necesario realizar ninguna tarea previa para la generación del árbol de decisión. Utilizando la misma metodología de la sección anterior (Hold-Out de 1/3) se obtiene el árbol de decisión de la figura 2

Tras analizar el árbol se ha comprobado que 2 de los atributos (**PRE19** y **PRE32**) no se han escogido para la construcción del árbol de decisión, por tanto, posteriormente se ha vuelto a generar a partir del algoritmo *J48* otro árbol de decisión, pero esta vez eliminando del conjunto de datos dichos atributos. La matriz de confusión obtenida en con todos los atributos se muestra en la tabla 3. La tasa de acierto resultante de dicha clasificación ha sido del 75 %.

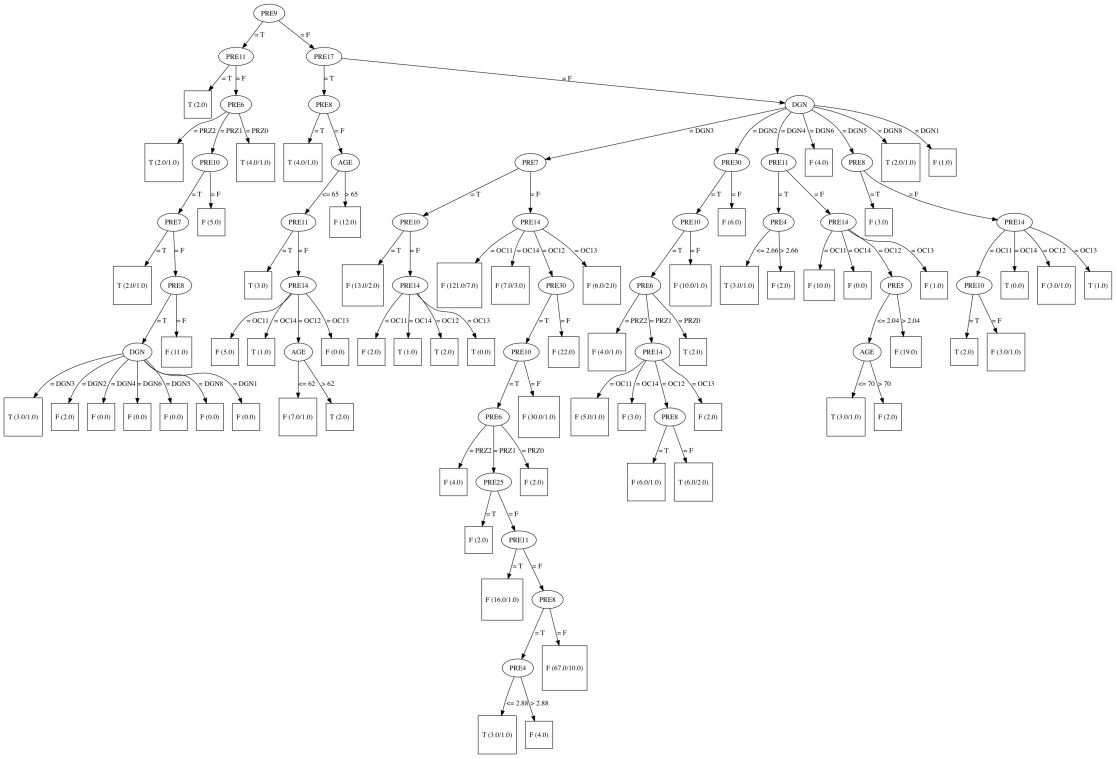

Figura 2: *Árbol de decisión generado a partir del algoritmo* J48

		Valor Real		
		Positivo	Negativo	p_j
Valor Predicho	Positivo	4	18	0,1375
	Negativo	22	116	0,8625
	π_j	0,1625	0,8375	$N = 160$

Tabla 2: *Matriz de confusión del conjunto de datos entrenado por el algoritmo* J48

El árbol resultante de aplicar el algoritmo *J48* eliminando los atributos **PRE19** y **PRE32** se muestra en la figura 3. Además, en la tabla 3 se muestra la matriz de confusión, cuyos resultados son equivalentes al caso anterior. A pesar de ello el árbol de decisión presenta una estructura diferente.

La razón por la cual sucede esto es debido a que la implementación del algoritmo *J48* implementada en *Weka*[3] utiliza una medida ligeramente diferente de **Ganancia de Información** con respecto de la básica. La principal diferencia subyace en que la ganancia basada en ratio divide el resultado de la Ganancia en Información entre la entropía del atributo, lo cual produce una variación de resultados dependiendo del número de atributos que tenga el conjunto de datos. Esto se ilustra de forma más clara en la ecuación (1)

$$GainRatio(Class, Attribute) = \frac{Gain(Class, Attribute)}{H(Attribute)} = \frac{H(Class) - H(Class|Attribute)}{H(Attribute)} \tag{1}$$

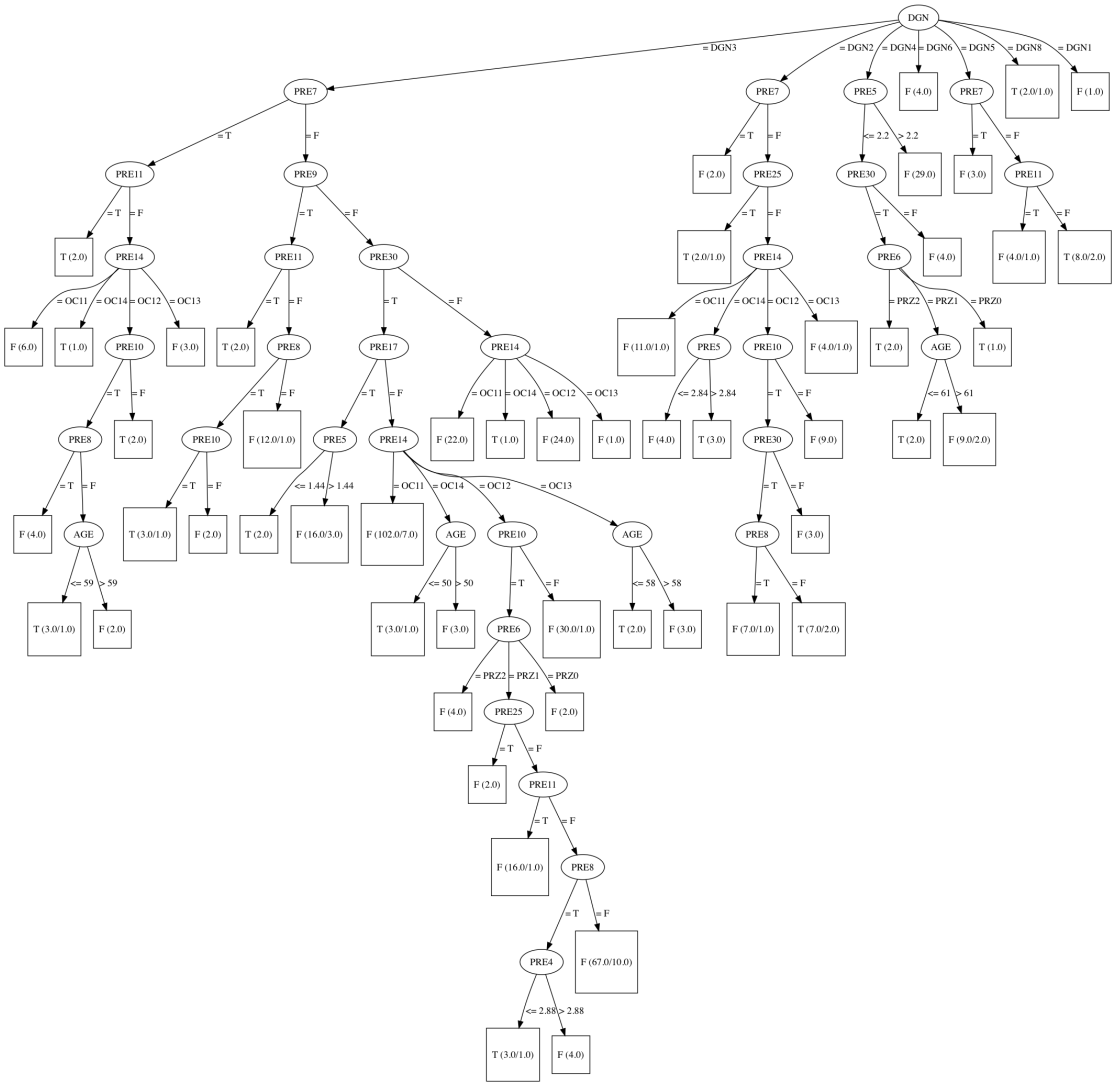

Figura 3: *Árbol de decisión generado a partir del algoritmo* J48 *tras eliminar los atributos no utilizados previamente.*

		Valor Real		
		Positivo	Negativo	p_j
Valor Predicho	Positivo	4	18	0,1375
	Negativo	22	116	0,8625
	π_j	0,1625	0,8375	$N = 160$

Tabla 3: *Matriz de confusión del conjunto de datos entrenado por el algoritmo* J48*tras eliminar los atributos no utilizados previamente.*

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?

La técnica seguida por el algoritmo *J48* ofrece mejores resultados que la aplicada en el apartado 2 con el algoritmo *ID3* debido a que en dicho caso la discretización se realiza *A priori*. Esto quiere decir que además de no poder apoyarse en los valores de la clase de destino para la clasificación, este atributo tan solo podrá ser utilizado una vez para la clasificación.

Por contra, el algoritmo *J48* realiza una técnica que aplica particionamiento binario y recursivo utilizando la *Ganancia de Información* como medida de bondad. La técnica de particionamiento recursivo de atributos numéricos permite una mejor clasificación ya que tan solo se produce cuando se maximiza la ganancia de información para dicha partición binaria.

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

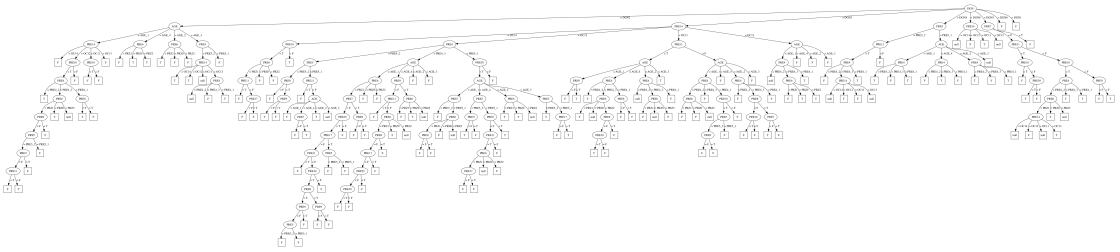

Figura 4: *Árbol de decisión generado a partir del algoritmo* ID3 *discretizado según las particiones del apartado 3*

La matriz de confusión resultante de dicha clasificación se muestra en la tabla 4. Nótese que se han utilizado **160 instancias** pero el árbol de decisión ha dejado sin clasificar 2 de ellas, por lo que la tasa de acierto global es del 77,5 %.

		Valor Real		
		Positivo	Negativo	p_j
Valor Predicho	Positivo	4	13	0,1075
	Negativo	21	120	0,8924
	π_j	0,1582	0,8417	$N = 158$

Tabla 4: *Matriz de confusión del conjunto de datos discretizado a partir de la división de 2 y después entrenado por el algoritmo* ID3

Referencias

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

[2] Taboada Rodero, I. J. treetograph. https://github.com/ismtabo/treetograph.

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

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