Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 12, Mes Diciembre, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 34-39
http://publicaciones.uci.cu

Tipo de artículo: Artículo original
Temática: Soluciones Informáttica
Recibido: 02/11/17 | Aceptado: 10/12/17 | Publicado: 17/12/17

Selección aleatoria de criterios de división

Random selection of division criteria

Alejandro Giubel Hernández Arbelo [1*], Mailyn Moreno Espino [2]

[1] Universidad Tecnológica de la Habana, CUJAE. aherandeza@ceis.cujae.edu.cu, my@ceis.cujae.edu.cu

[*] Autor para correspondencia: aherandeza@ceis.cujae.edu.cu

Resumen

La clasificación es una de las tareas más utilizadas de la minería de datos. Dentro de ella existen numerosos algoritmos como las redes neuronales, k vecinos más cercanos, clasificadores bayesianos entre otros. Uno de los más utilizados son los árboles de decisión, debido a su fácil comprensión de estructura jerárquica. Existen numerosos algoritmos generadores de árboles de decisión como el ID3, C4.5 y los CART. En este trabajo proponemos una variante del algoritmo ID3 que se basa en la selección aleatoria del criterio de división de los datos.

Palabras clave: clasificación; minería de datos; árboles de decisión; ID3; C4.5; CART; criterio de división.

Abstract

The classification is one of the most used tasks of data mining. Within it there are numerous algorithms such as neural networks, k nearest neighbors, Bayesian classifiers among others. One of the most used are decision trees, due to its easy understanding of hierarchical structure. There are numerous algorithms that generate decision trees such as ID3, C4.5 and CART. In this paper we propose a variant of the ID3 algorithm that is based on the random selection of the data division criterion.

Keywords: *classification; data mining; decision trees; ID3; C4.5; CART; division criterion.*

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 12, Mes Diciembre, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 34-39
http://publicaciones.uci.cu

Introducción

Uno de los puntos más importantes a la hora de la construcción de un árbol de decisión es la selección del atributo que se utilizará para dividir los datos[1]. Esta selección no se realiza de manera aleatoria, estudios han demostrado que mediante el uso de la estadística se puede seleccionar un atributo bajo un criterio matemáticamente respaldado, como el algoritmo ID3 que utiliza la Ganancia de Información como criterio de división de los datos[2, 3] o los CART [4]que hacen uso del Índice de Gini como criterio. Existen numerosos criterios para la selección del mejor atributo [5] por lo que no existe un criterio mejor que otro, ya que cada uno se comporta de maneras diferentes en diferentes entornos.

El siguiente trabajo muestra los resultados del algoritmo R-ID3, que basa su funcionamiento en la selección aleatoria de criterios de división para la selección del mejor atributo.

Materiales y métodos o Metodología computacional

Muchos criterios se basan en el uso de la entropía que es capaz de medir el nivel de desorden de un sistema, como la Ganancia de Información o la Proporción de Ganancia otros utilizan variantes estadísticas como el Índice de Gini o Chi-cuadrado.

Entre los criterios más conocidos se encuentran la Ganancia de Información[3], el Índice de Gini y la Proporción de Ganancia[3][6, 7]. El primero es utilizado por el algoritmo clásico de generación de árboles de decisión ID3[2, 3],el segundo es utilizado en algoritmos como los CART [4] y el ultimo se utiliza en las versiones mejoradas del ID3 como el C4.5 [3, 8] y el C5.0[3].

Algoritmo R-ID3.

El algoritmo ID3 es un algoritmo recursivo que, en cada llamada de este, el conjunto de datos cada vez va siendo menor hasta quedar completamente vacío, la cual es una de sus condiciones de parada. Podríamos resumirlo en que en cada llamada del ID3 el conjunto de datos es diferente, viéndolo en su totalidad.

En la figura 1 se evidencia que un criterio no siempre es mejor que los otros, es decir no existe un criterio único que siempre dé el mejor resultado. Apoyados en esta idea y que en cada iteración del algoritmo el conjunto de datos es diferente, podríamos utilizar un criterio diferente en cada iteración del algoritmo.

R-ID3 es una versión modificada del algoritmo clásico de generación de árboles de decisión ID3. Su única diferencia es que R-ID3 no utiliza un único criterio de división de los datos como en el caso del ID3 que utiliza la Ganancia de Información. R-ID3 utilizada tres criterios diferentes de división de los datos (Ganancia de Información, Índice de Gini y Proporción de Ganancia), seleccionando un criterio de manera aleatoria en cada iteración del algoritmo. Esto provoca que el mejor atributo por el cual dividir los datos no se seleccione siempre por el mismo criterio de división (aunque existe la posibilidad ya que la selección es aleatoria), acción que influye en la construcción del árbol resultante.

Seudocódigo.

Inputs: *R: a set of non- target attributes, C: the target* ***attribute****, S: training data.*

Output*: returns a decision tree*

Start

Initialize to empty tree;

If *S is empty then*

  Return *a single node failure value*

End If

If *S is made only for the values of the same target* ***then***

  Return *a single node of this value*

End if

If *R is empty* ***then***

  Return *a single node with value as the most common value of the target attribute values found in S*

End if

D ← the best attribute selected by a random selectedcriteria splitting (Information Gain, Gini Index, Gain Ratio)

{djj = 1, 2, ..., m} ← Attribute values of D

{Sjwith j = 1, 2, ..., m} ←The subsets of S respectively constituted of djrecords attribute value D

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 12, Mes Diciembre, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 34-39
http://publicaciones.uci.cu

Return *a tree whose root is D and the arcs are labeled by d1, d2, ...,dmand going to sub-trees ID3 (R-{D}, C, S1), ID3 (R-{D} C, S2), .., ID3 (R-{D}, C, Sm)*

End

Resultados y discusión

Descripción de las bases de datos utilizadas.

Todas las bases de datos utilizadas fueron descargadas de la *UCI machine learningrepository*. A cada base de datos se le realizó un pre-procesamiento que consistió en eliminar los atributos continuos y las filas con valores faltantes, debido a que el algoritmo utilizado fue el ID3.

Para los datos de prueba se utilizó el 30% de las bases de datos, los cuales fueron seleccionado de manera aleatoria, quedando un 70% para el entrenamiento del algoritmo.

Tabla. 1. Comparación entre ID3 y R-ID3en cuanto a su precisión.

Bases de datos	**Ganancia de información**	**Índice de Gini**	**Proporción de ganancia**	**R-ID3**
adult	77.7458	77.7826	77.9522	78.3576
breast	54.2168	51.8072	61.4457	63.8554
breast-w	88.2352	88.2352	88.2352	91.6666
bridge	65.625	65.625	65.625	65.625
car	90.5405	90.5405	89.9613	90.5405
chronickidneydisease full	63.4615	63.4615	78.8461	78.8461
cmc	90.2494	90.0226	90.2494	90.9297
credit-g	40	42.8571	44.0816	45.7142

crx	78.1094	79.6019	80.0995	80.0995

La tabla 1 muestra las precisiones obtenidas del algoritmo ID3 con los tres criterios de división y la precisión del algoritmo R-ID3.

El algoritmo R-ID3 fue ejecutado en 50 iteraciones, seleccionándose el mejor resultado de estas. Convirtiéndose esta en su principal desventaja, ya que si solo se ejecuta una sola vez las probabilidades de obtener un resultado mejor que los del ID3 se minimizan.

Conclusiones

Según los datos arrojados por las pruebas obtenidas podemos concluir que la construcción de un árbol de decisión con diferentes criterios de división puede proporcionar buenos resultados.

El algoritmo R-ID3 demuestra que existe una combinación de criterios de decisión que es capaz de dar mejores resultados que utilizar un solo criterio de división.

Referencias

[1] E. H. Jr., "Information Gain Versus Gain Ratio: A Study of Split Method Biases," 2001.

[2] J. R. QUINLAN, "Induction of Decision Trees," *Machine Learning,* 1986.

[3] A. M. Badr Hssina, Hanane Ezzikourl,Mohammed Erritali, "A comparative study of decision tree ID3 and C4.5," *International Journal of Advanced Computer Science and Applications*.

[4] R. A. Berk, "Classification and Regression Trees (CART)," pp. 1-66, 2008.

[5] Y.-S. SHIH, "Families of splitting criteria for classification trees," *Statistics and Computing,* vol. 9, pp. 309-315, 1999.

[6] V. B. Kishor Kumar Reddy, "A Survey on Issues of Decision Tree and Non-Decision Tree Algorithms," *International Journal of Artificial Intelligence and Applications for Smart Devices,* vol. 4, pp. 9-32, 2016.

[7] J. C. C. Fernando Berzal , Fernando Cuenca,Maríıa J. Martín-Bautista, "On the quest for easy-to-understand splitting rules," *Data & Knowledge Engineering,* vol. 44, pp. 31-48, 2003.

[8] S. L. Salzberg, "C4.5: Program for Macging Learning by J. Ross Quinlan.," vol. 16, pp. 235-240, 1994.