Balance entre confianza, soporte y comprensibilidad en la evolución de reglas de clasificación

Emiliano Carreño[†], Guillermo Leguizamón[†]

[†]Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC)
Departamento de Informática
Universidad Nacional de San Luis
Ejército de Los Andes 950 - Local 106
(D5700HHW) - San Luis - Argentina
Tel: (02652) 420823 / Fax: (02652) 430224
e-mail: {ecarreno, legui}@unsl.edu.ar

Resumen

En este artículo se propone un método para lograr un balance adecuado entre los parámetros de soporte, confianza y simplicidad al asignar los valores de fitness en la evolución de reglas de clasificación por medio de programación genética (PG). Un procedimiento adaptativo permite ajustar los valores de los parámetros de la técnica en orden a lograr tal balance. Este trabajo se enmarca dentro del campo de la minería de datos, poniendo especial atención en la extracción de conocimiento comprensible donde la técnica introducida juega un papel preponderante.

Palabras Claves: Minería de Datos; Reglas de Clasificación; Programación Genética; Ranking; Conocimiento Comprensible.

1. Introducción

La aplicación de programación genética para el descubrimiento de reglas de clasificación a partir de un conjunto de datos, presenta un inconveniente cuando cuando el tamaño de los árboles (*S-expresiones*) crece de manera importante. En tal caso, la complejidad del modelo obtenido hace casi imposible comprender el comportamiento del proceso subyacente generador de los datos. Así, si el modelo obtenido está formado por muchas reglas de alta complejidad, su entendimiento puede ser tan difícil de comprender como una red neuronal.

Por otra parte, los parámetros de soporte (o completitud) y confianza (o precisión) determinan la calidad de una hipótesis dada. Sin embargo, se debe dar un balance entre estos dos parámetros para que el modelo sea adecuado. Por ejemplo, una regla con una confianza de $0{,}5$ no aporta información acerca de la pertenencia de una instancia a una clase dada, tampoco sería de mucha ayuda una regla con un valor alto de confianza y poco soporte.

El enfoque propuesto en este artículo pretende establecer un balance adecuado entre la confianza, el soporte y la complejidad (relacionada directamente con la comprensibilidad) de una regla mediante la incorporación de un procedimiento adaptativo que hace una jerarquización de los individuos (ranking) basada en probabilidades y tomando en consideración los valores de soporte, confianza y comprensibilidad de los individuos en la población. Dicho procedimiento permite sesgar la búsqueda hacia regiones de hipótesis con una alta comprensibilidad y un balance adecuado entre soporte y confianza.

El problema de clasificación abordado en este trabajo incluye la predicción del precio de una vivienda (valor numérico dicretizado en tres intervalos) a partir de información sobre la zona en la cual se ubica (número de habitaciones de la vivienda, índice de criminalidad, etc.). El conjunto de datos denominado “Boston Housing”, el cual proviene del repositorio la Universidad de California en Irvine (UCI), cuenta con 13 atributos numéricos y uno binario con 506 instancias en total. Las instancias reflejan las condiciones de las viviendas en los suburbios de la ciudad de Boston.

2. Evolución de Reglas

En esta sección se presentan los aspectos más relevantes a tener en cuenta para la extracción de reglas de clasificación mediante el uso de PG en el contexto de minería de datos. La idea principal de PG es la evolución de programas de computadora (con estructura jerárquica de árbol) los cuales producen una solución para el problema en cuestión. Dados los conjuntos de funciones y terminales, un modelo (solución) se obtiene a partir del proceso evolutivo. El conjunto de funciones puede contener operadores ariméticos y lógicos, entre otros. El conjunto de terminales contiene las variables del problema así como tambien la constante aleatoria efímera $\Re$, representando números aleatorios con un determinado rango y precisión. La habilidad de los individuos en la población para resolver el problema en cuestion se mide mediante la función de adaptación (función

de fitness). Luego de la creación de la población inicial, el algoritmo se ejecuta generación por generación hasta que se satisface el criterio de tertminación, después de lo cual se selecciona la mejor solución encontrada. En cada generación se evalua la aptitud (fitness) de cada individuo, seleccionándose de forma probabilística los mejores individuos en la población en base a algún método de selección para luego aplicar los operadores de reproducción, crosover y mutación (cada uno en base a una determinada probabilidad). Para una descripción más detallada del paradigma de la programación genética se puede consultar [1].

Las reglas evolucionadas en este trabajo son del tipo SI ⟨*condición*⟩ ENTONCES ⟨*consecuente*⟩ (IF-THEN). El antecedente de la regla está formado por combinaciones lógicas de condiciones sobre los valores de los atributos predictores usando los conectivos lógicos AND, OR y NOT. Mientras que la parte del consecuente indica la clase a la cual se asigna una determinada instancia.

Para evaluar la calidad de la reglas se emplean las medidas de confianza (precision) y soporte (alcance). La confianza se calcula como el cociente entre el número de instancias a las cuales la regla se aplica y predice correctamente entre el número de instancias a las cuales la regla se aplica. Esto es, la confianza da la probabilidad de que la regla clasifique correctamente una instancia a la cual se aplica. El soporte (o completitud) se calcula como el cociente del número de instancias a las cuales la regla se aplica y predice correctamente entre el número total de instancias de la clase.

En la presente propuesta se considera el uso de programación genética para la evolución de reglas de clasificación donde cada individuo se representa mediante un árbol que codifica únicamente el antecedente de la regla. Esto se debe a que se ejecuta el programa genético tantas veces como clases distintas existan. El conjunto de funciones incluye los operadores lógicos AND, OR y NOT juntamente con el operador de igualdad vinculando cada atributo con alguna clase. El operador de igualdad se aplica sobre los atributos discretizados durante la evolución de las reglas. El conjunto de terminales está formado por la selección de un subconjunto de los 13 atributos predictores mas la constante aleatoria efímera $\Re$.

La selección de los atributos junto con el proceso de discretización y la selección de los conjuntos de entrenamiento y test se lleva a cabo empleando la herramienta de minería de datos WEKA (Waikato Environment for Knowledge Analisys). La comprensibilidad es de suma importancia dentro del contexto de minería de datos. Así pues, el objetivo principal de este trabajo es la obtención de reglas comprensibles al usuario. Si bien la compresibilidad es un concepto muy subjetivo, aquí se la mide por medio de la complejidad sintáctica de las reglas. Dicha complejidad se obtiene contando el número de nodos en el árbol sintáctico. Al evolucionar reglas de clasificación debe tenerse en cuenta que las mejores soluciones en cuanto al soporte y confianza pueden no encontrarse en las regiones del espacio de búsqueda donde las hipótesis tengan la comprensibilidad deseada. Por ello hay que alcanzar un consenso entre la comprensibilidad por un lado y una combinación del soporte y la confianza por el otro.

3. Algoritmo Propuesto

En la literatura se encuentran publicados trabajos dentro del contexto evolutivo donde en el proceso de descubrimiento de conocimiento se intenta conseguir reglas con alta capacidad predictiva, comprensibles e interesantes. En [2] se presenta un enfoque para el descubrimiento de reglas de predicción interesantes mediante el empleo de un algoritmo genético donde la función de adaptación (fitness) se divide en dos partes. Una parte mide el grado de interés de las reglas y la otra la exactitud predictiva. Por su parte en [3] se plantea el uso de programación genética para el descubrimiento de reglas comprensibles donde se incluye una penalización de la complejidad en la función de adaptación. Otras formas de lograr estos objetivos es mediante una matriz de confusión, o bien mediante un algoritmo genético con enfoque multiobjetivo [4].

La propuesta de este trabajo incluye la aplicación de un enfoque que trata de lograr el balance deseado entre compresibilidad y capacidad predictiva mediante el empleo de un algoritmo estocástico y adaptativo que forma un ranking de las soluciones candidatas considerando de forma probabilística los factores de soporte, confianza y complejidad (comprensibilidad) de las soluciones. Dicho ranking de soluciones puede ser alcanzado mediante algun algoritmo de ordenamiento (e.g., quicksort de Hoare) aplicando ciertos criterios de comparación basado en probabilidades que son ajustadas adaptativamente de acuerdo a una función que se retroalimenta del proceso de búsqueda. Las probabilidades son las siguientes:

1. P_Sop: es la probabilidad de usar la medida de soporte al comparar dos soluciones.
2. P_Conf: es la probabilidad de emplear la medida de confianza al realizar la comparación.
3. P_Long: es la probabilidad de emplear la medida de comprensibilidad al comparar dos individuos.

donde $P\_Sop + P\_Conf + P\_Long = 1$.

Para realizar el ordenamiento, la comparación de dos soluciones se lleva a cabo de forma excluyente de acuerdo a las medidas de soporte, confianza y longitud en base a las probabilidades antes mencionadas de la siguiente forma:

1. Se selecciona un número aleatorio $rnd \in [0, 1]$.
2. Si $0 \leq rnd < P\_Sop$ entonces la comparación se hace en base al soporte.
3. Si $P\_Sop \leq rnd < P\_Sop + P\_Conf$ la comparación se realiza de acuerdo a la confianza.

4. Si $P\_Sop + P\_Conf \leq rnd$ se hace la comparación de acuerdo a la medida de complejidad.

Este enfoque tiene ciertas similitudes con el método *Stochastic Ranking* de Runarsson y Yao [5] usado en el contexto de manejo de restricciones en algoritmos evolutivos a fin de establecer un balance entre las funciones de penalización y la función objetivo. Sin embargo, la técnica propuesta en el presente artículo posee las siguientes diferencias fundamentales (respecto al método de Runarsson et al.) que permiten balancear de forma directa tres factores importantes: soporte, confianza y longitud.

1. Las probabilidades P_Sop, P_Conf y P_Long tienen un significado totalmente diferente a cualquier probabilidad introducida en [5].

2. Se incorpora un procedimiento adaptativo para ajustar de forma adecuada los valores de P_Sop, P_Conf y P_Long.

3. Finalmente el método que se propone aquí no sólo intenta manejar una restricción blanda como lo es la complejidad (estructural) de las soluciones sino que también tiene como objetivo establecer un balance entre dos medidas que determinan el fitness de una solución (soporte y confianza). En [6] se plantea que la función de adaptación es la medida aritmética del soporte y la confianza de la regla representada en el cromosoma. Sin embargo, en este enfoque el fitness de un individuo no se establece directamente como una medida arimética de los valores de soporte y confianza sino que las probabilidades antes mencionadas intervienen probabilísticamente en el fitness de un individuo y por lo tanto en su posición en el ranking.

Luego, la probabilidad de que una hipótesis sea selecionada en el procedimiento de comparación está dada por la Eq.(1).

$$\begin{aligned} P(X) \;=\; & P(\, X.Sop > Y.Sop \,) \quad P\_Sop \\ & + P(\, X.Conf > Y.Conf \,) \quad P\_Conf \\ & + P(\, X.Long > Y.Long \,) \quad P\_Long \end{aligned} \tag{1}$$

Los valores de P_Sop, P_Conf y P_Long son ajustados de forma adaptativa de acuerdo a:

1. un parámetro de la técnica definiendo el límite de complejidad a partir del cual las soluciones son penalizadas.

2. valores de soporte, confianza y complejidad de las soluciones en la población actual o un subconjunto de ésta.

El criterio de selección de la mejor solución en cada generación está dado por la Eq.(2) la cual define la métrica F_β. El resultado de una ejecución se establece en consecuencia, como el individuo con el valor más alto de F_β encontrado en alguna generación.

$$F_\beta = \frac{(1+\beta^2)\ \textit{soporte} \ \cdot \textit{confianza}}{\beta^2 \ \cdot \textit{confianza} + \text{soporte}} \tag{2}$$

4. Resultados Preliminares y Consideraciones Finales

El método propuesto en este artículo intenta obtener un balance entre el soporte, confianza y longitud en la evolución de reglas de clasificación. El objetivo es sesgar la búsqueda hacia regiones del espacio de búsqueda con las características deseadas, i.e., alcanzar hipótesis comprensibles y con una alta calidad predictiva.

De acuerdo a los resultados preliminares se observa, en principio, que el enfoque propuesto logra mantener la búsqueda en regiones del espacio con hipótesis de complejidad estructural que permite lograr una comprensibilidad adecuada. En cuanto a los valores de soporte y confianza en general se consiguen resultados de mayor o igual calidad a los obtenidos sin el empleo del método propuesto. Esto es, la calidad de las soluciones no disminuye al disminuir la complejidad estrucctural aumentando la comprensibilidad del modelo obtenido.

El tiempo de ejecución necesario para la evolución de las reglas disminuye significativamente con el empleo del método. La reducción de la complejidad estructural de las hipótesis permite su más rápida evaluación permitiendo contrarestar cierta complejidad introducida por la incorporación del método.

Finalmente, cabe destacar que el enfoque propuesto en este artículo está en pleno desarrollo pudiéndose plantear mejoras en los siguientes aspectos:

1. Procedimiento adaptativo.

2. Criterio de designación del resultado.

3. Discretización de los atributos predictores.

Consecuentemente, los resultados preliminares sugieren la realización de trabajos desde el punto de vista teórico y experimental, los cuales pueden llevar a una mejora tanto de la compresibilidad de los modelos como de su calidad predictiva.

Referencias

[1] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[2] Edgar Noda, Alex A. Freitas, and Heitor S. Lopes. Discovering interesting prediction rules with a genetic algorithm. In Peter J. Angeline, Zbyszek Michalewicz, Marc Schoenauer, Xin Yao, and Ali Zalzala, editors, *Proceedings of the Congress on*

Evolutionary Computation, volume 2, pages 1322–1329, Mayflower Hotel, Washington D.C., USA, 6-9 1999. IEEE Press.

[3] Celia C. Bojarczuk, Heitor S. Lopes, and Alex A. Freitas. Genetic programming for knowledge discovery in chest-pain diagnosis. *IEEE Engineering in Medicine and Biology Magazine*, 19(4):38–44, July-August 2000.

[4] Kalyanmoy Deb and Deb Kalyanmoy. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 2001.

[5] T. Runarsson and X. Yao. Stochastic ranking for constrained evolutionary optimization, 2000.

[6] H. Orallo R. Quintana F. Ramírez. *Introducción a la Minería de Datos*. PEARSON EDUCACIÓN, S.A, Ribera del Loira, 28 28042 Madrid (España), 2004.