Evolución de Reglas de Clasificación para el Descubrimiento de Conocimiento Comprensible

Línea de Investigación: Agentes y Sistemas Inteligentes

Emiliano Carreño, Guillermo Leguizamón
Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC), Departamento de Informática, Universidad Nacional de San Luis, (D5700HHW)-San Luis - Argentina
(Tel: 2652-420823; fax: 2652-430224; email: {ecarreno, legui}@unsl.edu.ar).

Neal Wagner
Department of Mathematics & Computer Science, Augusta State University, 2500 Walton Way Augusta, GA 30904 USA (Tel: (706)667-4479; email: nwagner@aug.edu).

RESUMEN

Este trabajo, el cual se encuentra dentro del contexto de la minería de datos, propone un método para construir clasificadores basado en la evolución de reglas. El método, denominado REC (Rule Evolution for Classifiers), tiene tres características principales: 1) aplica programación genética (PG) para llevar a cabo una búsqueda en el espacio de potenciales soluciones, 2) un procedimiento permite sesgar la búsqueda hacia regiones de hipótesis comprensibles con alta calidad predictiva, 3) incluye una estrategia para la selección de un subconjunto óptimo de reglas (clasificador), a partir de las reglas obtenidas como resultado del proceso evolutivo. Se lleva a cabo un estudio comparativo entre este método y el algoritmo de inducción de reglas C5.0, para dos problemas de aplicación (conjuntos de datos). Los resultados experimentales muestran las ventajas de usar el método propuesto.

DESCRIPCIÓN DE TRABAJOS DE INVESTIGACIÓN EN EJECUCIÓN

La Figura 1 ilustra el método propuesto. La idea es evolucionar reglas de clasificación, sesgando la búsqueda hacia regiones de hipótesis comprensibles con alta calidad predictiva (características 1 y 2). Luego, una estrategia de selección construye el conjunto de reglas correspondiente al modelo final (clasificador) usando las mejores soluciones generadas durante el proceso evolutivo (característica 3).

La aplicación de PG para el descubrimiento de reglas de clasificación a partir de un conjunto de datos no es conveniente cuando el tamaño de los árboles (s-expressions) se incrementa significativamente. En tales casos, la complejidad del modelo obtenido hace que sea casi

imposible comprender el proceso generador de datos subyacente. Así, si se obtiene un modelo compuesto de muchas reglas de alta complejidad, éste puede ser tan difícil de comprender como una red neuronal compleja. Por otra parte, las medidas de soporte y precisión determinan la calidad predictiva de una hipótesis dada. No obstante, un modelo apropiado debería proveer un balance adecuado entre ambos parámetros. Por ejemplo, una regla con 0.5 de precisión no provee ninguna información sobre si una instancia pertenece o no a una determinada clase, sin embargo, una regla de alta precisión y bajo soporte tampoco es muy útil.

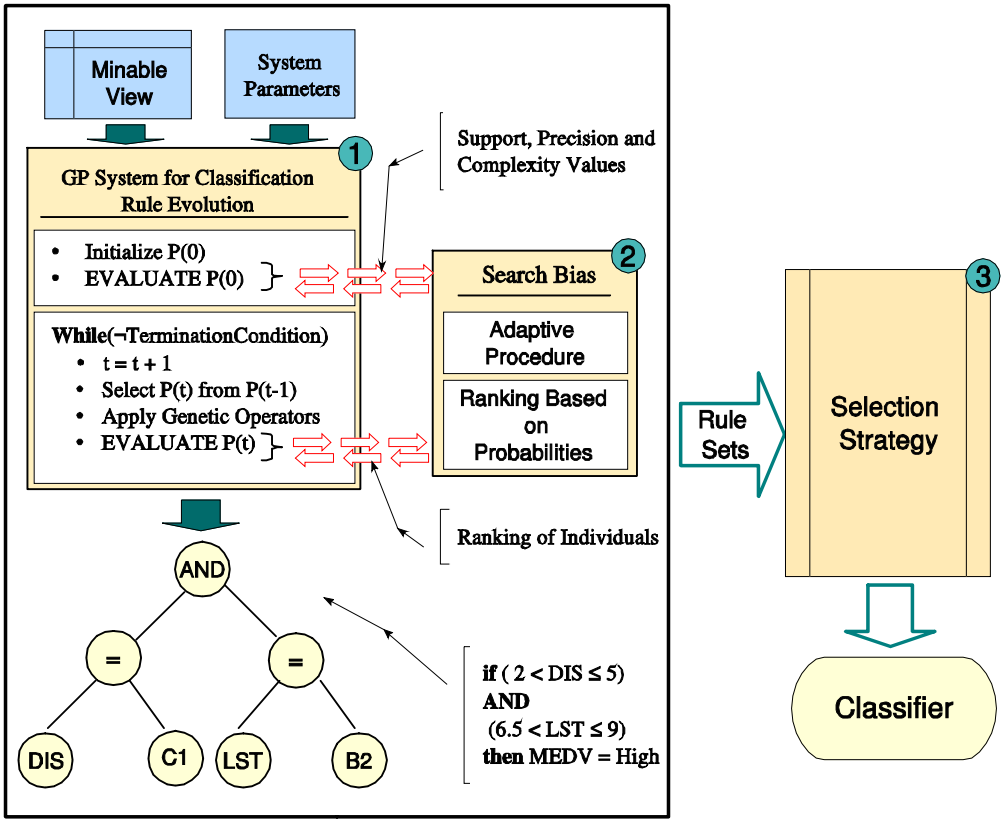

Figura 1: El método propuesto. 1) Sistema de PG para la evolución de reglas de clasificación. 2) Procedimiento para sesgar la búsqueda. 3) Estrategia de selección.

El enfoque propuesto en este trabajo intenta establecer un balance adecuado entre el soporte, la precisión y la complejidad (directamente relacionada con la comprensibilidad) de una regla mediante la incorporación de un procedimiento adaptativo el cual hace un ranking de las soluciones de forma probabilística, basándose en los valores calculados de soporte, precisión y comprensibilidad. Este procedimiento permite sesgar la búsqueda hacia regiones de hipótesis con alta comprensibilidad y un balance apropiado entre soporte y precisión. Para cada clase del atributo objetivo, se obtiene un conjunto de reglas como resultado de un proceso evolutivo. Luego, estos conjuntos se combinan mediante una estrategia para obtener el modelo final. La

idea es explotar las soluciones generadas durante el proceso evolutivo, seleccionando un conjunto de reglas óptimo (en cuanto a su calidad predictiva). En este sentido, puede ser que el conjunto de reglas óptimo no esté formado por las mejores soluciones encontradas durante el proceso evolutivo, sino por reglas que se complementen de forma adecuada. El resultado final es un clasificador expresado como un conjunto de reglas del tipo **if-then**.

En la literatura hay varios estudios donde el proceso de descubrimiento del conocimiento esta centrado en la obtención de reglas comprensibles, interesantes y con alta capacidad predictiva. Algunos ejemplos incluyen [1], [2], [3]. En [3], se presenta un enfoque para descubrir reglas de predicción interesantes mediante la aplicación de un algoritmo genético en el cual la función de adaptación (fitness function) esta dividida en dos partes. Una parte mide el grado de interés de las reglas, mientras que la otra mide su capacidad predictiva. En [1], se propone el uso de PG para el descubrimiento de reglas comprensibles, donde una penalidad para la complejidad se añade en la función adaptativa. En [2] esto también se logra por medio de la aplicación de un algoritmo genético con un enfoque multi-objetivo. En el presente trabajo proponemos un nuevo enfoque para sesgar la búsqueda hacia regiones de reglas comprensibles con alta calidad predictiva, en varios problemas de aplicación. Además, se introduce una estrategia para la construcción de clasificadores por medio de la selección de un subconjunto de reglas obtenidas como resultado del proceso evolutivo. Esta estrategia intenta obtener como modelo final, un subconjunto óptimo de reglas comprensibles con alta calidad predictiva.

En este trabajo se lleva a cabo un estudio comparativo del método propuesto contra **C5.0** [4], un algoritmo de inducción de reglas (del estado del arte) para la construcción de clasificadores. Este estudio analiza principalmente la calidad predictiva y la comprensibilidad de los modelos obtenidos con estos dos métodos. También, se informa el tiempo de ejecución de cada enfoque. Los conjuntos de datos usados provienen del repositorio de la Universidad de California en Irvine (UCI) [5].

CONCLUSIONES

De acuerdo a los resultados obtenidos, se puede concluir que el enfoque propuesto es capaz de centrar la búsqueda sobre regiones donde las hipótesis tienen una complejidad estructural tal que permite su adecuada comprensión, sin disminuir su calidad predictiva. Las mejoras logradas respecto a **C5.0** son significativas. Los resultados del estudio comparativo muestran las ventajas de usar el método propuesto (**REC**), dado que los clasificadores obtenidos tienen mejor calidad

predictiva que aquellos obtenidos usando **C5.0**. En los experimentos llevados a cabo, **C5.0** es ejecutado usando una técnica de boosting para 1 y 10 pruebas (trials). Para todos los problemas, cuando **C5.0** es ejecutado con 1 trial, los clasificadores obtenidos con ambos métodos tienen una complejidad estructural similar. Por otra parte, al ejecutar **C5.0** con 10 trials (como lo recomiendan sus autores) la calidad predictiva de sus modelos mejora, pero su complejidad se incrementa notablemente y aún así su calidad predictiva se mantiene por debajo de la calidad predictiva de los modelos generados con el sistema **REC**.

Por otra parte, a **C5.0** le toma mucho menos tiempo construir el clasificador que al método propuesto. Este último resultado era de esperarse, dado que **REC** aplica un algoritmo evolutivo para llevar a cabo una búsqueda en el espacio de potenciales soluciones. No obstante, **REC** compensa esta desventaja al generar modelos con una mejor calidad predictiva.

Una ventaja adicional del sistema **REC** es que es posible establecer la complejidad y estructura del clasificador a ser construido. Esto se puede hacer por medio de dos parámetros que determinan la complejidad máxima de una regla y el máximo número de reglas en el modelo. Los resultados demuestran la alta performance del método propuesto para construir clasificadores comprensibles y con alta calidad predictiva.

POSIBLES LÍNEAS DE INVESTIGACIÓN A SEGUIR

Entre las líneas de investigación futuras se plantean las siguientes:

1. Aunque el tiempo de ejecución del sistema REC es razonable, se lo podría reducir por medio de enfoques paralelos. Por ejemplo, podría haber varios procesadores, cada uno evolucionando reglas para una clase diferente.
2. Se estudia la posibilidad de extender el presente trabajo para ser aplicado a problemas de predicción de series de tiempo (discretizando los atributos). En los problemas de predicción de series de tiempo se intenta predecir los valores futuros de cierta variable analizando un conjunto de sus valores pasados y aquellos de otras variables relacionadas. Por ejemplo, las agencias meteorológicas intentan predecir el valor futuro de la temperatura basándose en los valores pasados de esa temperatura y los valores pasados de otras variables relacionadas tales como la humedad relativa, la velocidad del viento, la dirección del viento, etc. Una extensión del sistema **REC** a problemas de predicción de series de tiempo requiere la discretización de todas las variables involucradas; de esta forma, el valor a predecir es un valor nominal ordenado. Siguiendo con el ejemplo anterior, los valores de temperatura (***T***)

se podrían convertir en una serie de valores ordenados {*Baja* ($\boldsymbol{T}$ <15), *Media* (15≤ $\boldsymbol{T}$ <25) y *Alta* ($\boldsymbol{T}$ ≥25)}.

3. En [10] Neal Wagner propone un modelo de programación genética dinámico específicamente adaptado para predicciones de series de tiempo en ambientes no estáticos. Éste incorpora métodos para adaptarse de forma automática a ambientes cambiantes así como también para retener conocimiento aprendido en ambientes encontrados previamente. En este sentido, una posible línea de investigación podría plantear la incorporación de dichos métodos en el enfoque propuesto en el presente trabajo, para obtener modelos comprensibles y de alta calidad predictiva en problemas de predicción de series de tiempo (con ambientes no estáticos).

REFERENCIAS

[1] C. C. Bojarczuk, H. S. Lopes, and A. A. Freitas, "Genetic programming for knowledge discovery in chest-pain diagnosis," IEEE Engineering in Medicine and Biology Magazine, vol. 19, no. 4, pp. 38–44, Jul.-Aug. 2000.

[Online]. Available: http://ieeexplore.ieee.org/iel5/51/18543/00853480.pdf

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

[3] E. Noda, A. A. Freitas, and H. S. Lopes, "Discovering interesting prediction rules with a genetic algorithm," in Proceedings of the Congress on Evolutionary Computation, P. J. Angeline, Z. Michalewicz, M. Schoenauer, X. Yao, and A. Zalzala, Eds., vol. 2. Mayflower

Hotel, Washington D.C., USA: IEEE Press, 6-9 1999, pp. 1322–1329.

[Online]. Available: citeseer.ist.psu.edu/noda99discovering.html

[4] R. Quinlan, "Rulequest research data mining tools." 2006. [Online]. Available:

http://www.rulequest.com/

[5] C. B. D.J. Newman, S. Hettich and C. Merz, "UCI repository of machine learning databases." 1998. [Online]. Available: http://www.ics.uci.edu/»mlearn/MLRepository.html

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

[7] O. L. Mangarasian, R. Setiono, and W. H. Wolberg, "Pattern-recognition via linear-programming: theory and application to medical diagnosis," in Large-Scale Numerical Optimization, 1990, T. F. Coleman and Y. Li, Eds. Philadelphia: SIAM, 1990, pp. 22–31.

[8] J. E. B. J, "Adaptive selection methods for genetic algorithms," in Proc. ICGA 1, 1985, pp. 101–111.

[9] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[10] Neal Wagner, Zbigniew Michalewicz, Moutaz Khouja, and Roy Mc Gregor, "Time Series Forecasting for Dynamic Environments: the DyFor Genetic Program Model" IEEE Transactions on Evolutionary Computation. 2005.