Doctoral Consortium: Algoritmos de aprendizaje de Sistemas Basados en Reglas Difusas Lingüísticas en entornos de alta dimensionalidad

Javier Cózar

Computing Systems Department - I^3A
University of Castilla-La Mancha - Spain
javier.cozar@uclm.es

- **Directores:** Luis de la Ossa Jiménez y José Antonio Gámez Martín.
- **Fecha de inicio de la tesis:** Abril de 2013.

1. Resumen

Este proyecto de tesis se centra en el ámbito de la minería de datos aplicado al aprendizaje de modelos predictivos, en concreto de los sistemas basados en reglas difusas. La idea de esta línea de investigación surge a raíz del trabajo desarrollado en [1], en el que se utiliza un algoritmo de búsqueda local como parte del sistema de aprendizaje, que permite llevar a cabo esta tarea de forma eficiente apoyándose en ciertas características de estos modelos. Aprovechando esta eficiencia, enfocaremos el diseño de los algoritmos de aprendizaje a problemas de alta dimensionalidad, los cuales son cada vez más comunes dada la tendencia actual en el crecimiento del volumen de los datos que generamos y almacenamos.

Los Sistemas Basados en Reglas Difusas (SBRDs) [2] son la aplicación más importante de la Teoría de Conjuntos Difusos. Estos modelos están constituidos por un conjunto de reglas difusas. Las principales ventajas de estos modelos son la capacidad de tratar con la imprecisión inherente a los datos del mundo real y su alto nivel de interpretabilidad.

Por un lado, la incertidumbre es un aspecto fundamental cuando se trata con datos del mundo real. Existen diferentes fuentes de ruido que hacen que los datos tengan un comportamiento impreciso (sensores de adquisición de datos con un margen de error asociado, por ejemplo). Estos sistemas tratan la incertidumbre de forma inherente, lo que los convierte en sistemas muy robustos.

Por otro lado, la interpretabilidad de un modelo también es un aspecto beneficioso desde dos puntos de vista diferentes. En primer lugar, un experto humano es capaz de introducir conocimiento en el modelo con gran facilidad, lo que permite aumentar la capacidad de predicción de estos modelos a la vez que se tiene un mayor control del comportamiento del sistema. En segundo lugar, una vez tenemos construido el modelo, un experto humano puede extraer información útil de él que le permita comprender el comportamiento del sistema que representa dicho modelo. Un ejemplo de la utilidad de este hecho se puede observar en [3],

en el que se aprende un SBRD para predecir el consumo eléctrico de diferentes edificios en la siguiente hora. La ventaja de utilizar estos sistemas, aparte de la capacidad de predicción en sí misma, es la capacidad de comprender el comportamiento del consumo eléctrico a lo largo del tiempo, de tal manera que si, por ejemplo, observamos una regla que indica que después de comer el consumo eléctrico se eleva, se puede estudiar el porqué y tratar de paliar el problema.

Un SBRD se descompone en varios elementos (ver figura 1). A continuación se describen cada uno de ellos, siguiendo el flujo de control cuando se procesa un ejemplo a la hora de realizar una predicción.

- Los datos de entrada suelen ser valores reales. Sin embargo, estos sistemas manejan internamente conjuntos difusos, por lo que en primer lugar es necesario **fuzzificar** los datos de entrada, es decir, convertir los valores reales en conjuntos difusos.
- Los valores fuzzificados son utilizados para disparar un conjunto de reglas difusas almacenadas en la **base de reglas**. Estas reglas están definidas en función de la **base de datos**, que contiene el dominio y los conjuntos difusos asociados a cada variable del problema.
- Las reglas disparadas serán utilizadas en el **motor de inferencia**. Éstas efectuarán una predicción individual y posteriormente se combinarán todas estas salidas para formar una única predicción global.
- Finalmente, la salida global puede ser un conjunto difuso (o no, dependiendo del sistema de reglas que utilice el SBRD). En caso de ser un conjunto difuso, se utilizará la interfaz de **defuzzificación** para transformar éste en un valor real. Este paso se realiza porque al igual que el valor de entrada se suele tratar de un valor real, también es común requerír un valor numérico en la salida.

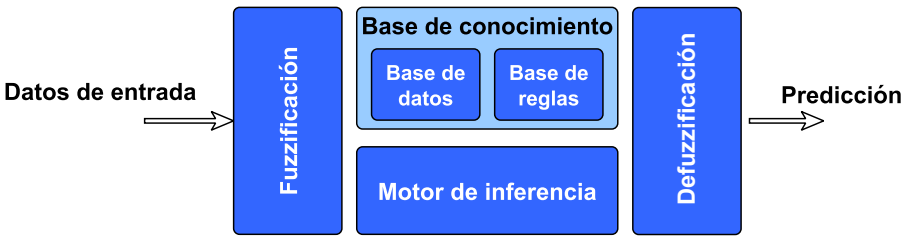

Figura 1: Módulos de un Sistema Basado en Reglas Difusas

La minería de datos se encarga de analizar un conjunto de datos para extraer un comportamiento o conjunto de patrones de él. Esta tarea se efectúa construyendo un modelo (en este caso, un SBRD). Existen diferentes algoritmos de aprendizaje de SBRDs. Principalmente, se pueden agrupar en tres conjuntos:

1. Derivan la base de datos, fijada previamente la base de reglas.
2. Derivan la base de reglas, fijada previamente la base de datos.
3. Generan simultáneamente la base de datos y de reglas.

En este proyecto se estudiarán y diseñarán algoritmos de aprendizaje de SBRDs, en principio, de dos de los tres grupos mencionados. En primer lugar nos centraremos en los del grupo 2, y posteriormente, nos centraremos en el aprendizaje de la base de datos junto con la base de reglas, aplicando en este último caso los algoritmos utilizados en el primer enfoque. Es posible que también se estudie el aprendizaje de modelos construyendo la base de conocimiento fijando el sistema de reglas (grupo 1) como estudio del comportamiento del sistema en función de ciertas parametrizaciones.

2. Metodología y plan de trabajo

Esta línea de investigación surge de [1], en el que se diseña un algoritmo de búsqueda local para aprender la base de reglas de un SBRD tipo Mamdani [4] fijada la base de datos previamente. Lo que se hace es, en primer lugar, generar un conjunto de reglas difusas candidatas (que potencialmente pueden formar la base de reglas), y posteriormente aplicar un algoritmo de búsqueda (un Hill Climbing) para seleccionar un subconjunto de estas reglas para formar la base de reglas final. La principal ventaja de utilizar el algoritmo de búsqueda local radica en la eficiencia, ya que aprovecha ciertas características de estos modelos para evitar realizar cálculos innecesarios.

En primer lugar se extenderá la idea presentada en este trabajo a los sistemas de reglas tipo TSK. Más concretamente a los sistemas TSK [5] de orden 0 y 1, en los que el consecuente de las reglas es un polinomio en función de las variables de entrada, de orden 0 y 1 respectivamente. Aunque el primer tipo posee un mayor nivel de interpretabilidad, los sistemas tipo TSK son más precisos. Por ello, dependiendo del problema a tratar, se utilizará uno u otro tipo de reglas, siempre buscando un equilibrio entre precisión e interpretabilidad.

Existen dos aspectos clave de los que adolece este algoritmo. Por un lado el enfoque utilizado en [1] sufre de un gran problema de escalabilidad, ya que uno de los pasos en los que se descompone el algoritmo (generación de reglas candidatas) tiene un orden de complejidad $\mathcal{O}(k^n)$, donde k es el número de conjuntos difusos por variable y n es el número de variables. Por otro lado, la expresividad e interpretabilidad de los modelos construidos se ve perjudicada por el hecho de fijar el número de antecedentes en cada regla difusa al máximo posible (número total de variables del problema). Por lo tanto, el siguiente paso será estudiar y diseñar una mejora para solventar ambos problemas.

Como se ha comentado previamente, [1] parte de una base de datos prefijada. Esto se hace derivándola en un primer paso a partir del conjunto de datos, dividiendo el dominio de cada variable equiespacialmente. Sin embargo, es muy dificil que este enfoque se ajuste al comportamiento real subyacente en el conjunto de datos, por lo que se limita en gran medida la precisión de los modelos

generados. Una alternativa que se estudiará es la de generar estos conjuntos difusos de manera más elaborada (en lugar de equifrecuencialmente). Otra vía que se estudiará es la de evolucionar conjuntamente el sistema de reglas y la base de datos. Este enfoque es muy común entre los denominados Sistemas Difusos Genéticos o *Genetic Fuzzy Systems* [6], que son SBRDs en los que el algoritmo empleado para la búsqueda del mejor conjunto de reglas (de entre las reglas candidatas) se realiza mediante un algoritmo genético, aplicando simultáneamente un proceso de refinamiento de la base de datos [7].

Como último objetivo, nos centraremos en el estudio de técnicas de aprendizaje online de modelos. En los esquemas tradicionales de aprendizaje automático, se parte de un conjunto de datos de entrenamiento a partir del cual se estima un modelo. No obstante, existen situaciones en las que la adquisición de estos datos de entrenamiento puede prolongarse hasta después de la construcción e implantación del modelo. El aprendizaje online consiste en mejorar la precisión del modelo obtenido con la llegada de nuevos datos, y es de esencial importancia cuando los sistemas se usan en entornos sujetos a cambios, es decir, cuando el sistema ha de ser adaptativo [8]. Las técnicas anteriormente descritas serán extendidas y adaptadas para su uso en entornos de aprendizaje online.

3. Relevancia

En esta tesis se pretende diseñar algoritmos de aprendizaje de diferentes modelos basados en reglas difusas lingüísticas. Esta variedad (sistemas Mamdani, TSK de orden 0 y 1, extensión a sistemas de aprendizaje online, ...) influye en que el campo de aplicación sea bastante extendido. Hoy en día el volumen de información que es manejado es inmenso, y los problemas del mundo real compuestos por pocas variables y/o instancias son cada vez más escasos. En este sentido, la posibilidad de aplicar estas técnicas a problemas de alta dimensionalidad para generar modelos compactos y altamente interpretables supone una gran aportación a la comunidad científica.

Los SBRDs han sido aplicados en diferentes campos, como problemas de series temporales, diagnóstico médico [9] o navegación autónoma de robots [10]. También existen trabajos que determinan niveles de equivalencia entre los sistemas basados en reglas difusas y las redes neuronales [11], modelos que actualmente están siendo muy utilizados [12]. Teniendo todo esto presente, pensamos en finalizar la tésis con un caso práctico en el que será aplicado uno de los métodos diseñados sobre un problema del mundo real, con la intención de conseguir modelos altamente interpretables y precisos.

Referencias

1. delaOssa, L., Gámez, J., Puerta, J.: Learning cooperative fuzzy rules using fast local search algorithms. In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2010). (2006) 2134–2141

2. Zadeh, L.: Outline of a new approach to the analysis of complex systems and decision processes. Systems, Man and Cybernetics, IEEE Transactions on (1) (1973) 28–44
3. Cózar, J., Vergara, G., Gámez, J.A., Soria-Olivas, E.: Comparing tsk-1 frbs against svr for electrical power prediction in buildings. In: International Joint Conference IFSA - EUSFLAT 2015. (2015) 880–887
4. Mamdani, E.: Application of fuzzy algorithms for control of simple dynamic plant. In: Proceedings of the Institution of Electrical Engineers. Volume 121., IET (1974) 1585–1588
5. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. Systems, Man and Cybernetics, IEEE Transactions on (1) (1985) 116–132
6. Cordón, O., Herrera, F., Gomide, F., Hoffman, F., Magdalena, L.: Ten years of genetic fuzzy systems: current framework and new trends. In: IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th. Volume 3., IEEE (2001) 1241–1246
7. Alcala-Fdez, J., Alcala, R., Herrera, F.: A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. Fuzzy Systems, IEEE Transactions on **19**(5) (2011) 857–872
8. Bermejo, P., Redondo, L., delaOssa, L., Rodríguez, D., Flores, J., Urea, C., Gámez, J., Puerta, J.: Design and simulation of a thermal comfort adaptive system based on fuzzy logic and on-line learning. Energy and Buildings **49** (2012) 367–379
9. Duckstein, L.e.a.: Fuzzy rule-based modeling with applications to geophysical, biological, and engineering systems. Volume 8. CRC press (1995)
10. Saffiotti, A.: The uses of fuzzy logic in autonomous robot navigation. Soft Computing **1**(4) (1997) 180–197
11. Jang, J.: Anfis: adaptive-network-based fuzzy inference system. Systems, Man and Cybernetics, IEEE Transactions on **23**(3) (1993) 665–685
12. Khaki, M., Yusoff, I., Islami, N.: Application of the artificial neural network and neuro-fuzzy system for assessment of groundwater quality. CLEAN–Soil, Air, Water **43**(4) (2015) 551–560