Minería de Datos utilizando Sistemas Inteligentes

Laura Lanzarini[1], Waldo Hasperué[2], Leonardo Corbalán[3], Sonia Formia[4], César Estrebou[5], Augusto Villa Monte[6], Germán Aquino[7], Marcela Jeréz[8]

Instituto de Investigación en Informática LIDI (III-LIDI)[9]
Facultad de Informática. UNLP

CONTEXTO

Esta presentación corresponde al Subproyecto "Sistemas Inteligentes" perteneciente al Proyecto "Procesamiento paralelo y distribuido. Fundamentos y aplicaciones en Sistemas Inteligentes y Tratamiento de imágenes y video" del Instituto de Investigación en Informática LIDI.

RESUMEN

Esta línea de investigación se centra en el estudio y desarrollo de estrategias adaptativas aplicables a la Minería de Datos.

Se han desarrollado distintas metaheurísticas que permiten obtener reglas y listas de clasificación, capaces de operar sobre datos numéricos y categóricos, a partir de datos estructurados etiquetados.

Se han analizado distintas técnicas no supervisadas para identificar las características más importantes de la deserción universitaria en base a la información de la UNRN.

Se está comenzando a trabajar con técnicas que operan con información no estructurada aplicables a la caracterización y clasificación de documentos.

Palabras clave: Minería de Datos, Minería de Textos, Reglas de Clasificación. Redes Neuronales. Hiper-rectángulos. Técnicas de Optimización.

1. INTRODUCCIÓN

En el Instituto de Investigación en Informática LIDI se está trabajando, desde hace varios años, en la resolución de problemas pertenecientes al área de Minería de Datos utilizando Sistemas Inteligentes.

Se han desarrollado diferentes técnicas de extracción de conocimiento utilizando diferentes herramientas inteligentes como las redes neuronales, la lógica difusa y las metaheurísticas de algoritmos genéticos y optimización por cúmulo de partículas.

Todas las técnicas desarrolladas trabajan sobre bases de datos estructuradas generando modelos en forma de reglas de clasificación o de asociación. En estos últimos meses se ha comenzado a trabajar con información no estructurada como el análisis semántico para la clasificación de documentos.

A continuación se detallan brevemente los avances realizados últimamente.

1.1. Obtención de reglas a partir de hiperrectángulos

Se ha desarrollado una técnica adaptativa, denominada CLUHR [5], que permite extraer conocimiento de grandes bases de datos a partir de un modelo dinámico capaz de adaptarse a los cambios de la información.

Esta técnica utiliza hiper-rectángulos, una poderosa forma de representación de datos,

[1] Profesor Titular DE. Facultad de Informática. UNLP
[2] Becario Post-doctoral (CONICET) – Jefe de Trabajos Prácticos - Facultad de Informática. UNLP
[3] Profesor Adjunto – Facultad de Informática – UNLP.
[4] Profesor Adjunto UNRN.
[5] Jefe de Trabajos Prácticos SD - Facultad de Informática. UNLP
[6] Becario CIN. Ayudante Diplomado - Facultad de Informática. UNLP
[7] Becario III-LIDI. Adscripto a cátedra - Facultad de Informática. UNLP
[8] Jefe de Trabajos Prácticos DE. Facultad de Ingeniería. UNPSJB.
[9] Calle 50 y 120 - 2do Piso, (1900) La Plata, Argentina, TE/Fax +(54) (221) 422-7707. http://weblidi.info.unlp.edu.ar

por su capacidad para describir de manera casi natural el subconjunto de datos al cual representa. Esto se debe a que los límites de cada hiper-rectángulo pueden ser utilizados como cláusulas en las reglas del tipo IF-THEN que resulten del proceso de extracción de conocimiento.

La estrategia propuesta comienza con un conjunto de hiper-rectángulos definidos a partir de los datos iniciales. Luego, en un proceso iterativo, se van eliminando superposiciones de acuerdo a los valores tomados por un conjunto de índices especialmente diseñados para este efecto. Como resultado de este proceso, los hiper-rectángulos cambian de tamaño o se dividen.

Este proceso de optimización continúa hasta minimizar (o anular) el volumen de intersección entre hiper-rectángulos de distintas clases. Finalmente se generan las reglas que resultan de los distintos hiper-rectángulos conseguidos.

La gran desventaja de CLUHR es que solo trabaja con atributos numéricos. Se ha desarrollado una mejora a esta técnica que incluye el tratamiento de atributos nominales. Esta mejora resulta en una técnica más poderosa denominada CLUIN [6].

Ambas técnicas, para llevar a cabo el armado del modelo de datos, deben tomar muchas decisiones. La decisión de que superposición resolver es decidida mediante el cálculo de los índices de superposición. Todas estas decisiones se configuran en forma previa al armado del modelo permitiendo que el algoritmo opere automáticamente.

Sin embargo, ambas técnicas son completamente flexibles permitiendo su ejecución de manera semi-automática o manual. Un experto puede intervenir en el proceso de obtención del modelo decidiendo ciertos aspectos claves del armado del mismo, logrando un resultado más acorde al problema presentado.

Otro aspecto importante de las técnicas desarrolladas es que presentan un comportamiento adaptativo, los nuevos datos que se ingresen al modelo de datos, causan la modificación de la estructura interna del modelo, evitando que el modelo no deba rehacerse nuevamente utilizando el conjunto de datos completo. Una vez actualizada la estructura interna del modelo se actualiza el conjunto de reglas existentes.

1.2. **Obtención de reglas a partir de técnicas de optimización**

Esta línea de investigación está centrada en la obtención de reglas de clasificación, del tipo IF-THEN, a partir de redes neuronales y técnicas de optimización.

En especial se estudian métodos de clustering y clasificación de patrones que permitan identificar, de una manera no supervisada, aquellos atributos relevantes para el problema. Dichos atributos serán especialmente considerados en el momento de construir el antecedente de la regla.

Este tipo de estrategias fueron utilizadas previamente en [11] para medir la relevancia de los términos más utilizados en un conjunto de e-mails. Sin embargo, las técnicas de agrupamiento no poseen la capacidad de seleccionar atributos. Dado que sólo operan con información numérica pueden utilizarse para obtener medidas de tendencia central para cada grupo a partir del conjunto de datos asociado. Esto último no implica el proceso de selección.

Como forma de identificar cuáles de los atributos son relevantes para la construcción de la regla, se investigan distintas variantes de optimización por cúmulos de partículas. Interesa especialmente el control adecuado de la velocidad ya que se relaciona directamente con la precisión de la respuesta obtenida.

Los resultados de esta investigación han dado lugar a una estrategia adaptativa capaz de generar una lista de reglas de clasificación reducida operando con atributos nominales y numéricos. La misma se basa en la combinación de una red neuronal SOM [7] y una técnica de optimización poblacional. Las reglas obtenidas se caracterizan por su

simplicidad y facilidad de interpretación dado que poseen pocos atributos en su antecedente.

Los resultados de esta investigación han sido publicados en [12]

1.3. Minería de Datos en Educación

La aplicación de técnicas de Minería de Datos en el ámbito educativo ha permitido caracterizar a los distintos actores que intervienen en los procesos de enseñanza-aprendizaje [10].

En el III-LIDI, se trabaja en este tema desde 2008. Las investigaciones realizadas han permitido evaluar la pertinencia y calidad del material desarrollado para un curso dado [4] [8]. También se estudiaron técnicas aplicables a la modelización del estudiante en lo referido a su proceso de aprendizaje [1] [9].

Actualmente, uno de los temas que más preocupa a las distintas unidades académicas es la deserción universitaria. Por tal motivo, se han estudiado distintas técnicas de modelización no supervisadas encontrando que las Redes Neuronales ofrecen una buena caracterización de los datos del problema. Se ha trabajado a partir de la información de los alumnos de la UNRN recolectados a través del sistema SIU-Guaraní exceptuando todos los datos sensibles del alumnado (DNI, Apellido y Nombre, nombre de los padres, etc).

Se realizó un preprocesamiento importante de la información que dio lugar a la vista minable. Luego a partir de distintos procesos de clustering basados en redes neuronales se identificaron las características más relevantes del problema llegando a la conclusión que la situación socio-económica del alumno tiene una fuerte incidencia en su permanencia en el ámbito universitario. Esto permitió demostrar también que las técnicas elegidas se adaptan al objetivo planteado.

Los resultados de esta investigación fueron presentados en [3]

1.4. Minería de Textos

La Minería de Textos posee los mismos objetivos generales que la Minería de Datos pero opera sobre colecciones de documentos de texto no estructurado. Las tareas que habitualmente se llevan a cabo pueden dividirse básicamente en las siguientes categorías: agrupamiento de documentos, categorización, clasificación y asociaciones de conceptos [2].

Esta línea de investigación tiene su eje central en el estudio y aplicación de distintos métodos de representación de documentos así como de distintas técnicas adaptativas aplicables en la resolución de problemas de agrupamiento y categorización.

Por el momento se ha utilizado el enfoque convencional basado en la generación de un diccionario de palabras y mediante distintas herramientas de visualización se han representado las relaciones más relevantes.

A futuro, interesa representar la relación semántica entre los términos que componen un mismo documento. Esto ayudaría a reforzar su importancia a la hora de buscar similitudes brindando mejores resultados.

2. TEMAS DE INVESTIGACIÓN Y DESARROLLO

- Investigación de nuevos índices de superposición de hiper-rectángulos con el objetivo de obtener mejores resultados ante problemas específicos.
- Estudio de técnicas de simplificación de modelos basados en reglas de clasificación.
- Estudio de distintas variantes de PSO capaces de controlar adecuadamente la velocidad con la que se mueven las partículas.
- Análisis de las limitaciones de PSO para operar sobre datos nominales. Identificación de las zonas más prometedoras del espacio de búsqueda utilizando SOM.

- Estudio de distintas técnicas de preprocesamiento aplicables a Minería de Textos.
- Estudio, análisis y comparación de diferentes técnicas de visualización.
- Estudio y desarrollo de métodos para la identificación de los atributos más relevantes de un conjunto de datos.
- Estudio de técnicas de agrupamiento aplicables a información numérica y categórica.

3. RESULTADOS OBTENIDOS/ ESPERADOS.

- Desarrollo de una técnica de extracción de conocimiento que opera sobre bases de datos con atributos numéricos y nominales.
- Desarrollo de técnicas capaces de adaptar su estructura interna ante la llegada de nuevos datos.
- Desarrollo e implementación de una representación para PSO de una regla de clasificación que incluya atributos numéricos y nominales.
- Desarrollo e implementación de un nuevo método capaz de obtener una lista de clasificación formada por un número reducido de reglas sencillas que operan tanto sobre atributos numéricos como nominales a partir de la combinación de una red SOM con una metaheurística basada en cúmulo de partículas.
- Desarrollo de una prueba de concepto que arroja información preliminar relevante respecto a la problemática del abandono.
- Descripción del perfil de los estudiantes aportando información útil en relación a su composición socio-económica y su permanencia en el ámbito universitario.
- Desarrollo de una herramienta para caracterización y clasificación de documentos.

4. FORMACIÓN DE RECURSOS HUMANOS

Dentro de los temas involucrados en esta línea de investigación se ha finalizado una tesis de doctorado y se están desarrollando actualmente 2 tesis de doctorado, 2 de maestría y al menos 3 tesinas de grado de Licenciatura. También participan en el desarrollo de las tareas becarios y pasantes del III-LIDI.

5. REFERENCIAS

[1] Arona, Huapaya, Lanzarini, Lizarralde. Lógica Difusa aplicada al Modelo del Estudiante de un Sistema Tutorial Inteligente. IV Congreso de Tecnología en Educación y Educación en Tecnología. TE&ET'09. Julio 2009. La Plata. Bs.As

[2] Berry, M.B., Kogan, J. (editors). Text Mining: Applications and Theory. John Wiley & Sons. 2010. ISBN 978-470-74982-1

[3] Formia S. Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios. Tesis de Especialista en Tecnología Informática aplicada en Educación. Dic 2012.

[4] Grossi, M.D., Lanzarini, L. Reglas de Predicción aplicables al Diseño de un Curso de Computación. III Congreso de Tecnología en Educación y Educación en Tecnología. TE&ET'08. Mayo 2008. Bahía Blanca.

[5] Hasperué, W., Lanzarini, L., De Guisti, A. Rule Extraction on Numeric Datasets Using Hyper-rectangles. Computer and Information Science. Vol. 5, No 4, pp. 116 131. 2012. http://dx.doi.org/10.5539/cis.v5n4p116

[6] Hasperué, W., Corbalan, L. CLUIN – A New Method for Extracting Rules for Large Databases. XIII Workshop de Agentes y Sistemas Inteligentes, XVIII Congreso Argentino de Ciencias de la Computación. Bahía Blanca. Argentina. Octubre 2012. Págs. 130-139.

[7] Kohonen, T. Self-Organizing Maps. 2nd Edition. Springer. ISSN 0720-678X (1997)

[8] Lanzarini, Denazis, Grossi Estrategias Inteligentes aplicables a un Sistema Educativo. X Workshop de Investigadores en Ciencias de la Computación (WICC 2008), Area Tecnología Informática Aplicada en Educación. Mayo de 2008. La Pampa

[9] Lanzarini, Huapaya. Diagnóstico Adaptativo del Estudiante en Sistemas Tutoriales Inteligentes. XI Workshop de Investigadores en Ciencias de la Computación (WICC 2009), Area Tecnología Informática Aplicada en Educación. Mayo de 2009. San Juan

[10] Romero C., Ventura S. Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, Volume 33, Issue 1, July 2007, Pages 135-146, ISSN 0957-4174.

[11] Villa Monte, A. Estrebou, C., Lanzarini, L. E-mail processing using data mining techniques. Publicado en el Libro *Computer Science & Technology Series – XVI Argentine Congress of Computer Science Selected Papers*, ISBN 978-950-34-0757-8. EDULP, Argentina , 2011.Págs. 109-120.

[12] Villa Monte, A., Ronchetti, F., Lanzarini, L., Jeréz, M. Obtención de reglas de clasificación usando SOM+PSO. XIII Workshop de Agentes y Sistemas Inteligentes". CACIC2012. ISBN: 978987-1648-34-4. Pág. 210-219. Bahía Blanca, Buenos Aires, Argentina, Octubre 2012.