Técnicas de clasificación aplicadas al estudio del rendimiento de ingresantes universitarios

Dapozo, Gladys; Porcel, Eduardo; López, María V., Greiner, Cristina
Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura
Universidad Nacional del Nordeste. 9 de Julio Nº 1449. CP 3400. Corrientes. Argentina.
TE: (03783) 423126 - (03783) 473930 Fax
{*gndapozo, eporcel, mvlopez, cgreiner*}*@exa.unne.edu.ar*

CONTEXTO

Este trabajo forma parte de las actividades definidas en el marco del proyecto "PI005/06: Análisis de Factores que inciden en el Rendimiento Académico y Desgranamiento de Alumnos de Carreras de la FaCENA", acreditado por la Secretaría General de Ciencia y Técnica de la Universidad Nacional del Nordeste (UNNE).

RESUMEN

Varios estudios han demostrado la incidencia de la formación en Matemática sobre el desempeño de los alumnos. Esta fortaleza resulta clave para el avance en las carreras que se ofrecen en la Facultad de Ciencias Exactas de la UNNE, dado que todas ellas tienen esta asignatura en el primer año. En este trabajo se propone realizar un estudio comparativo de técnicas de clasificación del campo de la minería de datos utilizando una herramienta de software libre. Se contrastará el rendimiento de las redes neuronales versus la aplicación de técnicas estadísticas convencionales. Se espera obtener resultados que permitan generar información útil para orientar políticas institucionales que contribuyan a mejorar los indicadores preocupantes de fracaso traducidos en deserción y bajo rendimiento de los alumnos del primer año de carreras universitarias.

Palabras clave: Minería de datos. Técnicas de clasificación. Herramienta de software libre. Rendimiento académico de alumnos universitarios.

INTRODUCCIÓN

La preocupación por el desempeño de los alumnos de primer año de carrera universitaria, que surge de los desfavorables indicadores de desgranamiento, abandono y rendimiento académico, ha llevado a las universidades del país a investigar sobre las causas que subyacen en esta problemática. La Universidad Nacional del Nordeste (UNNE) no es ajena a esta situación. En este sentido, ha realizado varios estudios con el objeto de aportar información que contribuya a configurar un cuadro de situación al interior de la institución [1].

La Facultad de Ciencias Exactas de la UNNE, con una matrícula de más de 6.000 alumnos, no escapa a la realidad descripta anteriormente. En esta Facultad, el Grupo de investigación en Matemática Aplicada a la Investigación Educativa, ha venido realizando desde el año 2000, trabajos de investigación que se han publicado en distintas revistas científicas y presentado, entre otros congresos, en las Reuniones Científicas que cada año organiza la Universidad, y que pueden consultarse en su página Web (http://www.unne.edu.ar/Web/cyt/presentacion.php).

Como subproyecto de este proyecto macro, se pretende abordar el análisis de los datos a través de diversas técnicas de minería de datos.

La minería de datos es un área de estudio científico con grandes expectativas para la comunidad investigadora, principalmente por las expectativas de transferencia a la sociedad. A pesar de la cantidad de publicaciones destacadas sobre la materia, queda por delante un campo fértil y prometedor con muchos retos en investigación [2].

Las técnicas que conforman el campo de la Minería de Datos buscan descubrir, en forma automática, el conocimiento contenido en la información almacenada en las bases de datos de las organizaciones. Por medio del análisis de datos, se pretende descubrir patrones, perfiles y tendencias. Es importante que estas técnicas sean las adecuadas al problema abordado. En este sentido, se pueden establecer dos grandes grupos de técnicas ó métodos analíticos: los métodos simbólicos y los métodos estadísticos [3].

Entre los métodos simbólicos se incluyen a las Redes Neuronales, Algoritmos Genéticos, Reglas de Asociación, Lógica Difusa, entre otros. Estos derivan del campo de la Inteligencia Artificial.

Los métodos estadísticos están constituidos por las técnicas del Análisis Multivariante de Datos, tales como Regresión Lineal Múltiple, Regresión No Lineal, Regresión Logística, Análisis Discriminante, Árboles de Regresión, entre otras. Las técnicas de esta categoría, de alguna manera, constituyen la piedra basal de la Minería de Datos [3].

El modelo de Regresión Logística es un método lineal que intenta modelizar la probabilidad de ocurrencia de un fenómeno. La variable dependiente es categórica dicotómica o policotómica, a los efectos de facilitar la interpretación [3]. Esta técnica resulta adecuada cuando se pretende hacer una clasificación basada en las características de los datos. Una ventaja adicional de esta técnica es que no requiere la normalidad estricta de los datos. Además, muchos estudios han evidenciado otras características que hacen de la regresión logística una buena herramienta para la categorización [4].

Una red neuronal es un sistema informático reticular (de inspiración neuronal) que aprende de la experiencia mediante la auto-modificación de sus conexiones [5], [6].

Las redes neuronales son modelos computacionales inspirados en las características neurofisiológicas del cerebro humano y están formadas por un gran número de neuronas dispuestas en varias capas e interconectadas entre sí mediante conexiones con pesos. Una neurona sobre un conjunto de nodos N es una tripleta (X, f, Y), donde X es un subconjunto de N, Y es un único nodo de N y f es una función neuronal que calcula un valor de salida para Y basado en una combinación de los valores de los componentes de X, es decir $y = f(\sum_{x_i \in N} w_i x_i)$. Los pesos w_i pueden ser positivos o negativos, reproduciendo el carácter excitador o inhibitorio de la sinapsis de las neuronas. Las redes neuronales usan un proceso de aprendizaje por analogía donde los pesos de las conexiones son ajustados para reproducir un conjunto de datos representativo del problema a aprender. Las redes neuronales constituyen herramientas analíticas que permiten examinar los datos con el objeto de descubrir y modelar las relaciones funcionales existentes entre las variables. Pueden comportarse como técnicas de aproximación o de clasificación universales [7].

Como antecedentes de aplicación de la técnica de redes neuronales en el ámbito de educación pueden mencionarse los trabajos de González [8], Salgueiro et al [9], Borracci y Arribalzaga [10].

El papel del software libre (SL) en la universidad es un fenómeno cuyas dimensiones éticas y sociales pueden transformar el marco académico, haciéndolo más democrático, participativo y viable en términos de recursos. A nivel académico, el SL refleja mucho mejor los valores tradicionales de la investigación universitaria desde su propia definición de “libre”: libertad para analizar cómo trabaja un programa y adaptarlo a nuestras necesidades, libertad para mejorar un programa y compartir con otros las adaptaciones, beneficiando así a toda la comunidad [11]. En la línea del software libre, está disponible el software Weka (*Waikato Environment for Knowledge Analysis)* que se encuentra de manera gratuita en Internet y contiene múltiples algoritmos para la aplicación de técnicas supervisadas y no supervisadas [12]. Es un producto con mayor orientación a las técnicas provenientes de la Inteligencia Artificial (IA) y de fuerte impacto en el contexto académico [13].

Este trabajo tiene los siguientes objetivos:

a) Realizar un estudio comparativo del rendimiento de redes neuronales, en concreto redes multi-capa de propagación hacia atrás (perceptrón multicapa), con modelos estadísticos convencionales, tales como regresión logística, en problemas de clasificación de una variable cualitativa de dos categorías de clasificación.
b) Analizar la eficiencia predictiva de estas técnicas aplicadas al estudio particular del rendimiento académico de alumnos del primer año de universidad.

METODOLOGIA

Para llevar a cabo este trabajo se requieren dos fases bien diferenciadas: la preparación de los datos y el análisis propiamente dicho.

La metodología dentro de la *fase de preparación* de los datos no está estandarizada, existen varias versiones según los autores. Para este trabajo se seleccionó la descripta en [14] que está compuesta por tres etapas principales, las cuales se dividen en otras subetapas de propósito específico:

Etapa 1. Selección y actualización de los datos

Se dispone de los datos de los alumnos provenientes del denominado Sistema de Ingreso de Alumnos de la UNNE. Este sistema incluye un formulario, en el cual los aspirantes a ingresar a la universidad hacen constar, además de sus datos de identificación personal, sus principales antecedentes sociodemográficos tales como edad, sexo, estado civil, lugar de procedencia y de residencia y sus antecedentes educacionales tales como tipo de título y de colegio de nivel medio del cual provienen, así como el nivel educativo alcanzado por los padres y el tipo de actividad económica y categoría ocupacional de los mismos.

Etapa 2. Preprocesado de Datos

- Integración de Datos: Los datos presentan formatos diferentes en los diferentes períodos de tiempo, debido a modificaciones del instrumento de recolección, de manera que deben ser sometidos a un proceso de integración y unificación de conceptos.
- Reconocimiento y Limpieza de Datos: El objetivo es reducir el ruido y las inconsistencias.

Etapa 3. Selección de Características

- Transformación de Datos: Implica la transformación del tipo de algunos atributos, en caso que fuera necesario, teniendo presente que convertir el tipo de un atributo a otro puede cambiar la semántica de dicho atributo.
- Reducción de Datos: Se eliminan características redundantes.

En la fase de *análisis*, se realizará la aplicación de las técnicas de clasificación.

Para la generación del modelo de datos se considerará la información de los años 2004 y 2005, cruzada con la información correspondiente al desempeño académico, que será tomada de los registros correspondientes a los informes de regularidad al finalizar el dictado de cada materia y de los registros correspondientes a los exámenes finales.

Todas las carreras de la FaCENA tienen en el primer cuatrimestre del primer año una materia con contenidos matemáticos (principalmente Algebra), que es necesario regularizar para cursar materias del segundo cuatrimestre. Por tal motivo, el rendimiento académico se medirá mediante una variable dicotómica que toma el valor 1 (uno) si el alumno regularizó o aprobó dicha asignatura, durante el primer año de estudios, y 0 (cero) en caso contrario.

Se considerará para el análisis el modelo de datos utilizado en un estudio previo [15] en el cual se aplicó la técnica de regresión logística binaria, por pasos hacia adelante, con un nivel de significación $\alpha=0.05$. De este modelo resultaron relevantes las siguientes variables: año de ingreso, carrera, sexo, tenencia de mail, orientación vocacional recibida, nivel del título secundario,

dependencia del establecimiento, cobertura de obra social, y nivel educacional del padre y de la madre.

Utilizando el mismo conjunto de datos, se aplicará una red neuronal de tipo perceptrón multicapa con la arquitectura que resulte más adecuada a los fines propuestos.

Para el proceso de entrenamiento de la red se presentará un conjunto de patrones de entrada, constituido por las variables que definen el perfil académico de los alumnos mencionadas anteriormente, y su correspondiente valor de salida esperado. Se utilizará un algoritmo de aprendizaje supervisado, ajustándose los pesos de forma que al final de este proceso, una vez aprendida la relación, la red sea capaz de clasificar correctamente un nuevo patrón que se le presente, indicando si el alumno regularizará/aprobará o no Matemática durante el primer año de estudios.

Analizando los diferentes aspectos a considerar en una herramienta de minería de datos [16], a los efectos del propósito y las condiciones de este estudio se utilizará la herramienta *open source* WEKA. Se estudiarán las distintas técnicas disponibles en la herramienta seleccionada y se aplicarán aquellas que se consideren más apropiadas al problema en estudio.

LINEAS DE INVESTIGACIÓN/DESARROLLO

Las técnicas y metodologías empleadas en esta línea de investigación, están relacionados con los siguientes campos:

- Minería de datos – Técnicas de preproceso y de clasificación
- Metodologías basadas en software libre
- Integración de datos (datawarehouse)
- Análisis estadístico
- Educación. Causas de bajo rendimiento académico en estudiantes universitarios

RESULTADOS ESPERADOS/OBTENIDOS

En esta línea de trabajo se han obtenido los siguientes resultados:

- Se han aplicado técnicas de integración de datos con la metodología de datawarehouse para mantener actualizado un repositorio con toda la información sistematizada existente en la unidad académica respecto del desempeño de los alumnos [17].
- Se ha realizado un estudio sobre técnicas de preprocesamiento que permitirá contar con información confiable y de mayor completitud [18].
- Se obtuvieron resultados preliminares sobre el perfil socioeconómico de los alumnos y su relación con el desempeño académico en el primer año de las distintas carreras de la Facena, utilizando técnicas de estadísticas clásicas [15].

Como resultados esperados se pretende:

- Profundizar el estudio de las técnicas de minería de datos para la explotación de datos, en particular los estudios de predicción y clasificación.
- Obtener información precisa sobre la controversia planteada en relación a qué modelos (estadísticos o neuronales) son más eficientes en la solución de problemas de clasificación en este modelo de datos orientado al rendimiento académico.
- Contribuir a brindar más información para orientar decisiones o acciones concretas destinadas a mejorar los preocupantes índices de desgranamiento, abandono y bajo rendimiento de los alumnos en el primer año de universidad.

FORMACIÓN DE RECURSOS HUMANOS

En esta línea de investigación se encuentra en curso el desarrollo del Trabajo Final de Aplicación de una alumna de la Licenciatura en Sistemas de Información de la UNNE.

BIBLIOGRAFIA

[1] Foio, Socorro. "El perfil socioeconómico de los ingresantes en la UNNE y su relación con la deserción en el primer año, la retención y el rendimiento académico" en http://www.unne.edu.ar/Web/estadistica/temainteres/Texto/Inf_Ingres/inf_ingres.htm visualizado el 14/02/2007.

[2] Riquelme José, Ruiz Roberto, Gilbert karina, "Minería de Datos: Conceptos y Tendencias, http://cabrillo.lsi.uned.es:8080/aepia/Uploads/29/308.pdf

[3] Britos, P. Minería de Datos. 1º Ed. Buenos Aires: Nueva Librería. ISBN 987-1104-30-8. 2005

[4] García Jiménez, M. et al. La predicción del rendimiento académico: regresión lineal versus regresión logística. Psicothema. Vol.12. Suplem. 2. 248-252.

[5] Hectht-Nielsen, R. Neurocomputing. Addison-Wesley. Cal. 1990.

[6] Hertz, J. Krogh, A. y Palmer, R. Introduction tom the theory of neural computation. Addison-Wesley. Cal. 1991.

[7] Castillo, E.; Cobo, A.; Gutiérrez, J.M.; Pruneda, R.E. Introducción a las Redes Funcionales con Aplicaciones. Un Nuevo Paradigma Neuronal. Editorial Paraninfo S.A. Madrid. España. pp.5-8; 8-16; 21-24, 30-34, 53-100. 1999.

[8] González, D.S. Detección de alumnos de riesgo y medición de la eficiencia de centros escolares mediante redes neuronales. Biblioteca de Económicas y Empresariales. Servicios de Internet. Universidad Complutense de Madrid. 1999.

[9] Salgueiro, F.; Costa, G.; Cánepa, S.; Lage,F.; Kraus, G.; Figueroa, N.; Cataldi; Z. Redes Neuronales para predecir la aptitud del alumno y sugerir acciones. WICC2006

[10] Borracci, R. A.; Arribalzaga , E. B. Aplicación de análisis de conglomerados y redes neuronales artificiales para la clasificación y selección de candidatos a residencias médicas. Educación Médica Vol 8 Nº 1. ISSN 1575-1813. Barcelona. 2005.

[11] Bustamante Donas, Javier. El software libre y la universidad. http://www.libroblanco.com/html/modules.php?op=modload&name=News&file=article&sid=164&mode=thread&order=0&thold=0.

[12] Machine Learning Project at the Department of Computer Science of The University of Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/

[13] Witten, IH and Frank, E: "Data Mining: Practical Machine Learning Tools and Techniques", 2nd Edition. Morgan Kaufmann, 2005

[14] González Sánchez, G. et al. Preprocesamiento de bases de datos masivas y multi-imensionales en minería de uso web para modelar usuarios. Universidad de Girona, España. En: http://eia.udg.es/~gustavog/esp/publicaciones/cedi2005_gustavo_sonia_published.pdf

[15] Porcel, E., Dapozo, G., López, María V. "Perfil socioeconómico y análisis del rendimiento académico en el primer año de los alumnos ingresantes de la Facultad de Ciencias Exactas de la UNNE". Enviado para su publicación a la Revista FaCENA en diciembre de 2007.

[16] Britos P, García-Martínez R., Selección de herramientas de explotación de datos. Una propuesta metodológica. http://www.itba.edu.ar/capis/rtis/rtis-6-2/seleccion-de-herramientas.pdf

[17] Dapozo, G., Porcel, E. "Metodología de integración de datos para apoyar el seguimiento y análisis del rendimiento académico de los alumnos de la FACENA". Comunicaciones Científicas y Tecnológicas de la UNNE 2005. http://www.unne.edu.ar/Web/cyt/com2005/8-Exactas/E-032.pdf.

[18] Dapozo G., Porcel E., López M. V., Bogado, V., "Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios". WICC 2007. ISBN 978-950-763-075-0. Universidad Nacional de la Patagonia San Juan Bosco. Trelew. Chubut. 2007.