Técnicas de Clasificación aplicadas al rendimiento académico

Myriam Herrera[1], María Inés Lund[2], Susana Beatriz Ruiz[1], Estela Liliana Torres[1], Lilian Adriana Mallea[3], María Gema Romagnano[2]

[1]Departamento de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan
[2]Instituto de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan
[3]Departamento de Matemática, Facultad de Filosofía, Humanidades y Artes, Universidad Nacional de San Juan

mherrera, mlund, mromagnano@{iinfo.unsj.edu.ar}

Resumen

En muchas investigaciones se tiene necesidad de identificar cuáles son las características que diferencian unos grupos de sujetos u objetos respecto de otros, para así poder realizar predicciones. El análisis de conglomerados y el análisis discriminante, son técnicas que algunos autores ubican entre las más potentes para aplicar en investigaciones sociales, permiten clasificar sujetos u objetos a partir de características similares.

Estas dos técnicas se pueden diferenciar por la manera de extraer conocimiento útil escondido en esos datos. El Análisis Discriminante cuenta con grupos de datos conocidos, con observaciones de unidades de pertenencia desconocida inicialmente y tiene que ser determinada a través del análisis de los datos. Este tipo de problemas de clasificación es referido como reconocimiento de patrones asistido o aprendizaje supervisado; en terminología estadística cae bajo el título de Análisis Discriminante.

Por otro lado, hay problemas de clasificación donde los grupos son desconocidos a priori y el principal propósito del análisis es determinar los grupos a partir de los propios datos, de modo que las unidades dentro del mismo grupo sean, en algún sentido, más similares u homogéneas que aquellas que pertenecen a grupos diferentes. Este tipo de problema de clasificación es referido como reconocimiento de patrón no supervisado o conocimiento sin guía, y, en terminología estadística cae bajo el título de Análisis de Conglomerados.

En este proyecto se aplicarán ambas técnicas o una combinación de ellas o una nueva técnica para analizar lo que llamamos rendimiento académico universitario. Se puede afirmar que, en general, un indicador directo de la calidad de la enseñanza es el rendimiento académico, medido a través del nivel alcanzado por los estudiantes. Vista la importancia del tema en este proyecto se determinarán las principales variables que influyen en el rendimiento como así también tipologías básicas de grupos, obtenidos de los alumnos universitarios tanto de la Facultad de Ciencias Exactas como de los alumnos de matemática de la Facultad de Filosofía de la UNSJ.

.

Palabras clave: Clasificación, Rendimiento, Calidad Universitaria

Contexto

Este proyecto se encuentra en un estado inicial, ya que recientemente ha sido aprobado por evaluación externa, es de carácter bi-anual y financiado por la UNSJ. Se encuentra inserto en el marco de las líneas de investigación de los Gabinetes Estadística e Ingeniería de Software del Instituto de Informática de la FCEFN de la UNSJ.

Además se vincula a cátedras de las carreras de Licenciatura en Ciencias de la Computación y Licenciatura en Sistemas de Información, que se dictan en la Institución.

Introducción

Los datos se han convertido en un recurso crítico en muchas organizaciones y por lo tanto, el acceso eficiente a estos, el compartirlos, extraer información de los mismos y hacer uso de la información extraída se transforma en una imperiosa necesidad. La necesidad de comprender conjuntos grandes y complejos de datos, ricos en información, es común a todos los campos de los negocios, ciencia, ingeniería. [1, 3, 11, 14, 15, 16]

Como resultado hay muchos esfuerzos no sólo para integrar varias fuentes de datos dispersos a través de sitios diferentes, sino también es importante la información extraída de esas bases de datos en bajo la forma de patrones y tendencias.

El proyecto se basa en dos principales ejes de investigación y aplicación:

Data Mining [3, 11] analiza conjuntos de datos para encontrar relaciones y resúmenes de datos útiles para el propietario de los datos. Estas relaciones y resúmenes derivados a través del ejercicio del Data Mining se refieren a modelos y patrones.

La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos. Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento.

El Reconocimiento de Patrones tiene como objetivo la **clasificación** de objetos dentro de un número de categorías o clases [7, 8, 12]. Dependiendo de la aplicación estos objetos pueden ser imágenes, señales o cualquier tipo de medidas que necesitan ser clasificadas. Esas medidas se llaman patrones.

Las medidas usadas para la clasificación de objetos o patrones son conocidas como características. El conjunto de todas las características forman el vector que identifica únicamente a un patrón (objeto).

Líneas de Investigación, Desarrollo e Innovación

En muchas de las investigaciones, independientemente del área de conocimiento, es habitual tener la necesidad de identificar cuáles son las características que diferencian unos grupos de sujetos u objetos respecto de otros, para así poder realizar predicciones futuras. Tanto el análisis de conglomerados como el análisis discriminante son técnicas que nos permiten clasificar sujetos u objetos a partir de características similares. La diferencia fundamental entre ambas pruebas es el momento del establecimiento de los grupos. En el análisis discriminante (AD) [2, 10, 13] el investigador conoce a priori a qué grupo pertenece cada sujeto u objeto; en

cambio, en el análisis de conglomerados los grupos o clúster se determinan y configuran a posteriori, es decir, una vez estudiadas y analizadas las agrupaciones.

El análisis discriminante es la prueba estadística apropiada para seleccionar qué variables independientes o predictivas permiten diferenciar grupos y cuántas de estas variables son necesarias para alcanzar la mejor clasificación posible. Además permite cuantificar su poder de discriminación en la relación de pertenencia de un sujeto u objeto a un grupo u otro. Por ello esta técnica es considerada, además de una prueba de clasificación, una prueba de dependencia. De hecho, su propósito es similar al análisis de regresión logística; la diferencia radica en que solo admite variables cuantitativas.

En el presente proyecto utilizaremos estas técnicas en el ámbito educativo como es el estudio del rendimiento estudiantil y la identificación de las variables que mejor lo predicen, a partir de las calificaciones de materias que clasifican al alumnado en grupos bien diferenciados.

Mediante un análisis discriminante se puede establecer el poder explicativo y discriminatorio de las características que diferencian a los alumnos según su rendimiento. Además del rendimiento se tendrán en cuenta en el estudio una serie de variables independientes como, por ejemplo, variables de carácter socioeconómico, variables académicas referentes a la preparación en el nivel secundario y variables actitudinales en relación con la variable dependiente que clasifica a los sujetos según el rendimiento obtenido [4, 9].

Según las características analizadas, a través de la descripción del grado de relación existente entre el conjunto de variables, se puede encontrar la frontera que separa los grupos.

Como resultado obtendremos una regla de clasificación que podrá ser utilizada en el pronóstico de adscripción al grupo de rendimiento establecido para nuevos estudiantes..

Resultados y Objetivos

Este proyecto es bi-anual y está iniciándose en 2016, con lo cual no tenemos actualmente resultados.

Objetivo General:

- Determinación de una función discriminante que explique la influencia de un conjunto de variables en el rendimiento académico de alumnos universitarios.

Objetivos Específicos:

- Determinar las Variables influyentes en el rendimiento académico de alumnos universitarios.
- Identificar qué variables tienen mayor poder de discriminación y de predicción en la clasificación de sujetos
- Determinar si existen diferencias significativas entre los “perfiles” de un conjunto de variables de dos o más grupos definidos a priori.
- Establecer un procedimiento, función discriminante, para clasificar a un individuo a partir de los valores de un conjunto de variables.
- Evaluar la exactitud de la clasificación mediante la regla de decisión que asigne un objeto nuevo a uno de los grupos prefijados con un cierto grado de riesgo

Formación de Recursos Humanos

El equipo de investigación está formado por docentes investigadores de dos facultades de la UNSJ, y las unidades de observación serán los datos de

alumnos de Licenciatura en Ciencias de la Computación y Licenciatura en Sistemas de Información del Departamento de Informática y Licenciatura en Matemática, del Departamento de Matemática.

Se espera sumar alumnos tesistas de grado y posgrado (maestría y doctorado), interesados en esta línea de investigación.

Referencias

1. ATO, M.; LÓPEZ, J.A. (1996): Análisis estadístico para datos categóricos. Madrid: Editorial Síntesis
2. BENZÉCRI, Jean Paul (1976): L´Analyse des données, T.I La taxonomie T.II L´Analyse des correspondances. Dunod. París.
3. CHENGKAI Li, «CSE4334/5334 Data Mining», University of Texas at Arlington, 2015.
4. DIAZ, M., PEIO, A., ARIAS, J., ESCUDERO, T., RODRIGUEZ, S., VIDAL, G. J. (2002). Evaluación del Rendimiento Académico en la Enseñanza Superior. Comparación de resultados entre alumnos procedentes de la LOGSE y del COU. En: Revista de Investigación Educativa, 2 (20), 357-383.
5. DIDAY, Edwin, y LECHEVALLIER, Yves (1991): Symbolic Numeric data analysis and learning, Versailles, September 18-20. INRIA, Nova Science Publishers Inc. New York.
6. DIDAY, Edwin (1992): Analyse des données et classification automatique numerique et symbolique. Seminario Internacional de Estadística en Euskadi. Volumen 27. EUSTAT, Euskal. Estatistika Erakundea/ Instituto Vasco de Estadística.
7. EUSTAT, Vitoria-Gasteitz. (1997): Análisis de datos simbólicos. Ed. IRICE, Rosario.
8. FERNÁNDEZ AGUIRRE, Karmele: IV International Meeting of Multidimensional Data Analysis (NGUS'97), Bilbao, September 10-12, 1997. Universidad del País Vasco, Bilbao.
9. GARBANZO VARGAS, Giuselle María (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la Educación Superior Pública. Revista Educación, 31(1), 43-63, ISSN: 0376-7082.
10. GONZALEZ LOPEZ, Ignacio (2004). Realización de un Análisis discriminante explicativo del rendimiento académico en la Universidad. Revista Investigación Operativa, vol 22, nº1, 43-59. Universidad de Córdoba.
11. HAN, Jiawei y KAMBER, Micheline, Data Mining. Concepts and Techniques, 2.a ed. Morgan Kaufmann, 2006.
12. LEBART, Ludovic; MORINEAU, Alain, y PIRON, Marie (1995): Statistique exploratoire multidimensionnelle. Dunod. París.
13. TORRADO FONSECA, Mercedes; BERLONGA-SILVENTE, Vanesa (2013). Análisis Discriminante mediante SPSS. REIRE (Revista d'Innovació i Recerca en Educació).
14. WEB. «Data mining made faster: New method eases analysis of "multidimensional" information», ScienceDaily. https://www.sciencedaily.com/releases/2010/07/100722075013.htm. [Accedido: 15-feb-2016].

15. WEB. «Big Data, for better or worse: 90% of world's data generated over last two years», *ScienceDaily*. https://www.sciencedaily.com/releases/2013/05/130522085217.htm. [Accedido: 15-feb-2016].
16. WEB «The Four V's of Big Data», *IBM Big Data & Analytics Hub*. http://www.ibmbigdatahub.com/infographic/four-vs-big-data. [Accedido: 15-feb-2016].