Técnicas de Minería de Datos como alternativa a las Técnicas Estadísticas de Discriminación y Clasificación Multivariadas Clásicas

Maria Paula DIESER[(1)], María Cristina MARTÍN[(2)], Erica SCHLAPS, Vanina Celeste BOLAÑO, Lorena CAVERO, María de los Ángeles IRRIBARRA Claudina SOLARO, Laura WAGNER, Diamela TITIONIK

Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa
Avenida Uruguay 151, (02954) 42-5166
[(1)] pauladieser@exactas.unlpam.edu.ar, [(2)] maritamartin@exactas.unlpam.edu.ar

Resumen

En este trabajo se describe brevemente una de las líneas de investigación que se están llevando a cabo en el Departamento de Matemática de la Facultad de Ciencias Exactas y Naturales de la Universidad Nacional de La Pampa, en relación a Métodos Multivariados Discriminantes y de Clasificación, y su sensibilidad y fiabilidad en la aplicación a diferentes problemas reales o simulados.

Si bien el estudio puede centrarse en ciertos métodos que podrían entenderse como clásicos y de una esencia más estadística, es indudable que, en los últimos años, se ha producido un gran crecimiento en las capacidades de generar y recolectar datos. En estos enormes volúmenes de datos, existe gran cantidad de información a la que sería difícil, cuando no imposible, acceder mediante los métodos clásicos. Técnicas propias de la Minería de Datos, posibilitan el análisis de estas masas de datos, en búsqueda de patrones y predicciones, que permitan generar información útil a partir de ellos.

Se pretende, entonces, comparar las diferentes técnicas estadísticas clásicas con las propias de la Minería de Datos en las tareas de Discriminación y Clasificación, estableciendo similitudes y diferencias, y analizando las estimaciones que se obtienen con ellas al aplicarlas a problemas reales o simulados.

Palabras clave: discriminación**,** clasificación, minería de datos, árboles de clasificación, redes neuronales, análisis de *clusters*.

Contexto

Desde 2014 se vienen realizando actividades de investigación, en el ámbito de la Facultad de Ciencias Exactas y Naturales (FCEyN) de la Universidad Nacional de La Pampa (UNLPam), relacionadas con el estudio y aplicación de Métodos Multivariados Discriminantes y de Clasificación, con el propósito de establecer similitudes y diferencias, y analizar las estimaciones que se obtienen con ellos al aplicarlos efectivamente en el Análisis de Datos Multivariados. El Proyecto, acreditado y financiado por la Institución mencionada, cuenta también con la participación de estudiantes de postgrado de la Universidad Nacional de Asunción (UNA).

Entre los métodos estudiados en el marco del Proyecto, se encuentran algunos propios de la Minería de Datos como: Árboles de Clasificación, Redes Neuronales y el Análisis de *Clusters.* Se ha desarrollado la teoría sobre algunas de estas técnicas, y realizado prácticas en el

manejo de diferentes herramientas de *software* basadas en la filosofía *Open Source*. Asimismo, se pretende programar subrutinas necesarias para la aplicación de los métodos estudiados mediante el lenguaje de programación R, y adquirir destreza en dicha aplicación con datos provenientes, preferentemente, de las áreas de investigación en la FCEyN de la UNLPam y la UNA y, caso sea necesario, realizar las simulaciones que permitan efectuar las comparaciones. En este sentido, el desarrollo del Proyecto, demanda una continua interacción con otros investigadores de la Institución, actuando en beneficio de sus avances.

Introducción

La característica principal del Análisis de Datos Multivariados (ADM) está en considerar un conjunto de *n* objetos sobre los que se observan los valores de *p* variables. El conjunto de objetos puede ser la totalidad o una muestra de un conjunto más grande. Las variables pueden ser cuantitativas (continuas o discretas) o cualitativas (nominales u ordinales), y a su vez podrían ser un subconjunto de un grupo más grande. Dada la situación planteada en cada caso, se busca estudiarlos con diferentes propósitos: simplificación de estructuras, clasificación, agrupación de variables, análisis de interdependencia y de dependencia, construcción y prueba de hipótesis, entre otras.

En el problema de discriminación y clasificación se dispone de un conjunto amplio de elementos que pueden provenir de dos o más poblaciones distintas. En cada elemento se han observado variables cuya distribución conjunta no necesariamente es conocida. Se desea diferenciar poblaciones o clasificar un nuevo elemento, con valores conocidos de sus variables, en poblaciones predefinidas. Dependiendo del tipo y del número, como así también de la distribución conjunta de las variables, existen diferentes enfoques para encarar los problemas de discriminación y clasificación.

El Análisis Lineal Discriminante (ALD) propuesto por Fisher (1936), está basado en la Distribución Normal Multivariada de las variables consideradas y es óptimo bajo este supuesto (Welch, 1939), siendo además necesaria la igualdad de matrices de covarianzas de los grupos considerados, aunque Anderson (1972) mostró una función discriminante para el caso donde no se cumple dicho supuesto. La técnica consiste en dividir el espacio muestral en subespacios mediante hiperplanos que permiten separar lo mejor posible los grupos en estudio; para ello, a partir de la matriz de datos se construye una función (lineal) que será usada para discriminar y clasificar las unidades experimentales. A menudo los datos no son normales por lo que es necesario transformar las variables para aplicar el método.

En aquellas situaciones en que las variables predictoras no están distribuidas normalmente y en las que algunas o todas esas variables son discretas o categóricas, se puede utilizar otro método de discriminación conocido como Regresión Logística (RL). Esta técnica, a menudo se usa para modelar la probabilidad de que una unidad experimental caiga en un grupo particular, con base en la información medida en la propia unidad. La técnica fue sugerida por Cornfield (1962), Cox (1966) y Day & Kerridge (1967) para una variable con respuesta binaria; Anderson (1972) propuso el modelo de Regresión Logística Multinomial.

Los métodos clásicos antes descriptos, que utilizan modelos estadísticos como habitualmente estamos acostumbrados a

formular, funcionan adecuadamente bajo los supuestos mencionados. Sin embargo, factores como el avance tecnológico vinculado al gran poder de procesamiento de las computadoras y su bajo costo de almacenamiento, han generado un enorme crecimiento en nuestras capacidades de generar y recolectar datos. Estas grandes masas de datos, en general provenientes de sistemas de información, tienen características particulares (alta cantidad de variables de diversos tipos, datos ausentes, ruido, entre otras) que las convierte en difícilmente tratables, cuando no imposible, mediante los métodos clásicos anteriores. Sin embargo, existen otras técnicas, propias de la Minería de Datos (MD), que permiten analizar estas masas de datos y generar infomación novedosa y útil a partir de ellos. Ejemplos de estas técnicas son los Árboles de Clasificación, las Redes Neuronales y el Análisis de *Clusters*.

Árboles de Clasificación

Los Árboles de Clasificación (AC) son métodos de aprendizaje automático y supervisado, útiles para encontrar estructuras en espacios de alta dimensionalidad y en problemas que mezclen datos categóricos y numéricos, que permiten construir modelos de clasificación de datos (Loh, 2011).

Estos modelos se obtienen particionando el espacio de datos en forma sucesiva y de a una variable por vez. El principio básico de la división jerárquica es obtener subcategorías del conjunto de observaciones de manera que éstas sean internamente lo más homogéneas posibles y se maximice la heterogeneidad entre los grupos. Existen diversos algoritmos, los más típicos se basan en un crecimiento de tipo *greedy*, enumerando toda posible partición de los datos disponibles en un cierto nodo para encontrar la mejor, entendida como aquélla que más disminuye cierta medida de impureza o diversidad al pasar del nodo padre a los nodos hijos. De esta manera se obtiene una serie de condiciones organizadas en forma jerárquica, a modo de árbol.

Shih (1999) y Martin (1997) ofrecen un estudio exhaustivo de diversos criterios de división existentes.

Las Redes Neuronales

Las Redes Neuronales (RN) permiten construir modelos para resolver problemas complejos de clasificación, regresión o agrupamiento, en los que puede haber interacciones no lineales entre variables de tipo numérico.

Una RN puede verse como un grafo dirigido constituido por nodos (elementos del proceso) y arcos entre ellos (sus interconexiones) (Hernández Orallo *et al.*, 2004). Estos nodos son unidades de procesamiento unitarias que reaccionan en paralelo. El nodo acepta una cantidad de información de entrada ponderada por pesos. Una función de activación, *f*, transforma las entradas recibidas en una información de salida y la envía hacia otra unidad (nodo) que la utiliza como información de entrada. El problema consiste, entonces, en encontrar la función *f* que mejor se ajusta a los datos observados y conduzca a un nodo final que permita la clasificación de las unidades experimentales según los factores que las caracterizan.

Análisis de *Clusters*

Los numerosos procedimientos que pueden encontrarse bajo la denominación Análisis de *Clusters* (AClu) están basados en la misma idea básica: dividir los individuos de una población en grupos (*clusters* o conglomerados) de manera tal que exista homogeneidad interna entre ellos y aislamiento externo entre los grupos. Para ello se debe procurar,

dependiendo del tipo de variables disponibles, una medida de disimilitud entre los individuos (euclidiana, Manhattan, Mahalanobis, son las más usuales, entre otras). Elegida ésta, se busca algún algoritmo que agrupe o clasifique los individuos con características similares (de Oliveira Bussab *et al.*, 1996; Kaufman & Rousseuw, 1990).

Las técnicas de agrupamiento pueden clasificarse en dos grupos: (a) Técnicas Jerárquicas, mediante las cuales los individuos son clasificados en grupos en diferentes etapas, de modo jerárquico, produciendo un árbol de clasificación (*Single*, *Complete, y Average Linkage*, Método del Centroide, entre otros); y (b) Técnicas no Jerárquicas, mediante las cuales los agrupamientos obtenidos en sucesivos pasos, producen una participación del conjunto original de individuos (Método de *k*- medias).

Líneas de Investigación

En el último año nos hemos enfocado en el desarrollo de la teoría y el conocimiento de diferentes Métodos Multivariados de Discriminación y Clasificación, tanto paramétricos como no paramétricos, entre los que consideramos algunos propios de MD que no utilizan un modelo estadístico como habitualmente estamos acostumbrados a formular.

En este contexto, una de las líneas de investigación es el análisis de las similitudes y diferencias entre las metodologías estudiadas, comparando las propiedades de las estimaciones cuando se aplican a datos de diferente naturaleza.

Se pretende, además, investigar nuevas técnicas de discriminación y clasificación emergentes, estableciendo equivalencias y diferencias con las ampliamente citadas.

Paralelamente se propone el desarrollo de prácticas en el manejo de diferentes programas estadísticos existentes para el procesamiento de datos, particularmente aquéllos basados en la filosofía *Open Source*, así como la programación de subrutinas necesarias para la aplicación de los diferentes métodos mediante el lenguaje de programación R.

Las técnicas estudiadas serán aplicadas sobre diferentes problemas del mundo real o, de ser necesario, a problemas simulados, con el fin último de comparar las estimaciones obtenidas y evaluar la calidad de las mismas.

Resultados y Objetivos

Hasta el momento nos hemos aproximado a los diferentes Métodos de Clasificación y Discriminación, objeto de estudio en el presente Proyecto (ALD, RL, y las que podrían enmarcarse dentro de MD como AC, RN y AClu). En este contexto, hemos realizado una detallada búsqueda bibliográfica lo que permitió desarrollar un estado del arte en relación a AL, RL, AC y AClu y, en particular, dos Tesis de Maestría de la UNA aplicando RL. Se espera completar el estado del arte de RN y avanzar en el estudio de otras técnicas emergentes.

Paralelamente se analizó la aplicación de las técnicas mencionadas en diferentes programas o paquetes de programas tales como Weka 3.6, RapidMiner Studio 6.0.008, y Rattle 3.3.0, utilizando conjuntos de datos clásicos. También se desarrollaron las rutinas necesarias para la aplicación de ALD y RL, mediante el lenguaje de programación R. Se prevé la programación de otras técnicas estudiadas o por estudiar usando el mismo lenguaje.

Se ha comenzado la identificación de diferentes bases de datos para realizar las experimentaciones correspondientes. Se pretende utilizar dichos datos para la aplicación de los métodos y su posterior comparación y análisis de fiabilidad.

Formación de Recursos Humanos

En el Proyecto, trabajan actualmente un Investigador formado (Directora), seis Investigadores en formación (tres de ellos de formación de base matemática y los restantes de áreas propias de las ciencias naturales), y dos Asistentes de Investigación (estudiantes avanzados de Licenciatura en Matemática cuya formación orientada se está realizando en temas de estadística). Todos ellos pertenecientes a la FCEyN de la UNLPam. Asimismo, participan del Proyecto, tres Tesistas de la UNA. Dos de estos trabajos de Tesis han sido defendidos satisfactoriamente (Sanabria, D.D., 2014; Vázquez, M.D., 2014).

Referencias

Anderson, J. A. (1972). Separate Sample Logistic Discrimination. *Biometrika*, 59(1):19–35.

Cornfield, J. (1962). Joint Dependence of the Risk of Coronary Heart Disease on Serum Cholesterol and Systolic Blood Pressure: A Discriminant Function Analysis. *Proceedings of the Federal American Society of Experimental Biology*, 21:58–61.

Cox, D. (1966). *Some Procedures Associated with the Logistic Qualitative Response Curve*. John Wiley & Sons, New York.

Day, N. E. & Kerridge, D. F. (1967). A General Maximum Likelihood Discriminant. *Biometrics*, 23(2):313–323.

de Oliveira Bussab, W., Miazaki, E. S., & de Andrade, D. (1996). *Introdução à Análise de Agrupamentos*. Asociación Brasilera de Estadística, San Pablo.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*. Pearson Prentice Hall, Madrid.

Kaufman, L. & Rousseuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Loh, W.-Y. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, 1(1):14–23.

Martin, J. K. (1997). An exact probability metric for decision tree splitting and stopping. *Machine Learning*, 28:257–291.

Sanabria, D. D. (2014). Determinación de variables bioclimáticas asociadas a la probabilidad de presencia de la especie de ave polinizadora Ermitaño Escamado (*Phaethornis eurynome*) en el Paraguay (tesis de maestría). Universidad Nacional de Asunción, Asunción, Paraguay.

Shih, Y.-S. (1999). Families of splitting criteria for classification trees. *Statistics and Computing*, 9(4):309–315.

Vázquez, M. D. (2014). Análisis del comportamiento, mediante la estimación y la predicción, de la morosidad de los socios de la Cooperativa Multiactiva de Ahorro, Crédito y Producción COOPEDUC - LTDA del Paraguay (tesis de maestría). Universidad Nacional de Asunción, Asunción, Paraguay.

Welch, B. L. (1939). Note on Discriminant Functions. *Biometrika*, 31:218–220.