Modelo de Decisión para la Validación de Métodos de Imputación Mediante la Utilización de Algoritmos de Minería de Datos

Carlos R. Primorac[1], Julio C. Acosta[1,2], David L. La Red Martínez[1]

[1] Facultad de Ciencias Exactas y Naturales y Agrimensura. Universidad Nacional del Nordeste
9 de Julio 1449, Corrientes (3400), Argentina

[2] Facultad de Ciencias Agrarias. Universidad Nacional del Nordeste
J. B. Cabral Nº 2131, Corrientes (3400), Argentina

carlosprimorac@gmail.com, julioaforever@hotmail.com, lrmdavid@exa.unne.edu.ar

Resumen

Muchos de los conjuntos de datos (data sets) existentes u obtenidos en investigaciones científicas contienen valores faltantes (MVs: Missing Values) y anomalías (outliers) asociados a procedimientos de entrada manuales deficientes, mediciones incorrectas o errores en los instrumentos de medición. En minería de datos (DM: Data Mining) estas imperfecciones pueden afectar negativamente la calidad del proceso de aprendizaje supervisado o el rendimiento de algoritmos de agrupamiento de datos. La imputación es una técnica para reemplazar MVs con valores sustituidos. Pocos estudios informan una evaluación global de los métodos existentes con el fin de proporcionar directrices para hacer la elección metodológica más apropiada en la práctica. El propósito general de este trabajo es determinar un modelo de decisión que permita encontrar los métodos de imputación más adecuados para completar información faltante en un conjunto de datos mediante la utilización de algoritmos de DM.

Palabras Clave: valores faltantes, imputación, minería de datos, modelo de decisión.

Contexto

La propuesta se inserta dentro de una de la líneas de trabajo del Grupo de Sistemas Operativos y TICs (Res. 725/10 C.D. - FaCENA) en el marco del Proyecto de Investigación "Incidencia de los perfiles de los alumnos en el rendimiento académico en Matemática del primer año de la Universidad", acreditado por la SGCyT - UNNE (PI: 16F002, Res. Nº 970/16 C.S.).

Diversos estudios y publicaciones abordan la evaluación de rendimiento académico utilizando técnicas de DM [1] [2] [3] [4] [5].

En este proyecto de investigación se propone evaluar el rendimiento académico de los estudiantes en las asignaturas Algebra de la carrera Licenciatura en Sistemas de Información (LSI) de la Facultad de Ciencias Exactas y Naturales y Agrimensura (FaCENA) y Matemática I de la carrera Ingeniería Agronómica (IA) de la Facultad de Ciencias Agrarias (FCA) de la de la Universidad Nacional del Nordeste (UNNE) utilizando técnicas de DM.

Para definir los perfiles de los estudiantes y determinar patrones que conduzcan al éxito o fracaso académico, se implementará un modelo que relaciona las calificaciones de los estudiantes con otras variables, tales como factores socioeconómicos, demográficos, actitudinales, entre otros; en base a lo cual se clasificaran los diferentes perfiles de alumnos.

Los modelos predictivos buscados, permitirán tomar acciones tendientes a evitar el fracaso académico, detectando los alumnos con perfil de riesgo de fracaso académico de manera temprana, a poco del inicio del cursado de las asignaturas; lo que permitirá concentrar en ellos los esfuerzos de tutorías y apoyos especiales.

Introducción

Históricamente, la noción de descubrir patrones ocultos en los datos ha recibido una variedad de denominaciones incluidos el de DM y descubrimiento del conocimiento (KDD: Knowledge Discovery in Databases). KDD, se refiere al proceso general de descubrir conocimiento útil a partir de los datos. La DM es una etapa dentro del proceso general de KDD que se refiere a los medios algorítmicos mediante los cuales se extraen y enumeran patrones a partir de los datos [6].

Muchos de los conjuntos de datos existentes u obtenidos en investigaciones científicas contienen MVs y anomalías (outliers) asociados a procedimientos de entrada manuales deficientes, mediciones incorrectas o errores en los instrumentos de medición. La presencia de estas imperfecciones generalmente requiere de una etapa de preprocesamiento en la cual, con el fin de que resulten útiles y suficientemente claros para el proceso de extracción de conocimiento, los datos se deben preparar y limpiar [7] [8] [9] [10] [11] [12].

En DM se pueden encontrar tres problemas principales asociados con MVs y outliers: i) pérdida de eficiencia, ii) complicaciones en la manipulación y análisis de los datos y iii) sesgo resultado de las diferencias entre valores faltantes y completos [10] [12] [13]. Estos afectar negativamente la calidad del proceso de aprendizaje supervisado o el rendimiento del algoritmo de agrupamiento de datos [11].

En la literatura se proponen dos enfoques generales para enfrentarse a los MVs. En el caso más simple, las instancias con MVs se omiten. Una segunda alternativa es utilizar técnicas de imputación y estimarlos utilizando los datos existentes [8] [10] [14] [15] [16] [17].

Tradicionalmente, el tratamiento de los MVs se realizaba antes del análisis de los datos mediante métodos diseñados “had hoc”. Algunas de estas estrategias consistían en trabajar con información completa, eliminando todos los casos con MVs en una o más variables (listwise-deletion), considerando los casos con valores disponibles en la variable de análisis y descartándolos cuando contienen MVs (pairwise-deletion) o sustituyendo MVs con el promedio de la variable considerada. Sin embargo, el sesgo introducido por estas técnicas ha hecho que sean fuertemente criticadas en la literatura [18].

La imputación es una técnica para reemplazar MVs con valores sustituidos. Una característica importante para una instancia en particular puede imputarse [15]. Estos métodos utilizan diferentes algoritmos que se pueden dividir en imputación simple (single-imputation) e imputación múltiple (MI: Multiple Imputation) y, en los últimos años, se ha propuesto el uso de algoritmos de aprendizaje automático (ML: Machine Learning) [9] [10] [11] [19] [14] [16] [18] [20] [21].

La selección de un método de imputación depende del conjunto de datos, el mecanismo de pérdida de datos y los patrones, el porcentaje de MVs y el desempeño de la técnica de imputación utilizada [15].

El mecanismo de pérdida de datos es un factor clave para decidir el método de imputación a utilizar. Rubin [22] definió tres mecanismos por los cuales se genera la perdida de datos: i) aleatoria (MAR: Missing at Random), completamente aleatoria (MCAR: Missing Completely at Random) y iii) no aleatoriamente (NMAR: Not Missing at Random).

El desempeño de un método de imputación, no solo depende de la cantidad de MVs, sino también de los patrones de pérdida. Existen diferentes patrones de MVs, algunos asociados con el registro y otros con el atributo. En el primer caso pueden ser simples, complejos, medios y mixto. En el segundo univariados, monótonos y arbitrarios [15]. Adicionalmente, el tamaño del conjunto

de datos y el porcentaje de MVs influye en la elección [16].

Diferentes técnicas de imputación funcionan bien sobre diferentes tipos de datos, algunas trabajan bien con enteros, otras únicamente con variables categóricas y algunas otras con datos combinados [12].

Finalmente, algoritmos de agrupamiento (no supervisados) y de clasificación (supervisados) se pueden adaptar para la imputación [14].

La mayoría de los artículos publicados en este campo se ocupan del desarrollo de nuevos métodos de imputación, sin embargo, pocos estudios informan una evaluación global de los métodos existentes con el fin de proporcionar directrices para hacer la elección metodológica más apropiada en la práctica [13].

1. Líneas de Investigación y Desarrollo

"Incidencia de los perfiles de los alumnos en el rendimiento académico en Matemática del primer año de la Universidad" es continuación de cuatro proyectos de investigación ejecutados desde el año 2004, oportunamente evaluados y acreditados en Comisión Externa; originados en la superpoblación y deserción de los alumnos en los cursos de trabajos prácticos de Algebra en el primer año de la Universidad. Se trabajará en la búsqueda de las variables que inciden en el rendimiento académico de los alumnos, para ejecutar las acciones que permitan evitar la deserción, corrigiendo las situaciones detectadas que la generan.

2. Resultado Esperados

El propósito general de este trabajo es determinar un modelo de decisión que permita encontrar los métodos de imputación más adecuados para completar información faltante en un conjunto de datos mediante la utilización de algoritmos de DM.

Se espera poder determinar una metodología para seleccionar los métodos de imputación de datos más adecuados para imputar cada variable del conjunto de datos. Se utilizarán como métodos de validación de los métodos de imputación algoritmos de DM de eficacia reconocida. El criterio de validación que se utilizará será el de mayor similitud entre los resultados de los procesos de minería antes de la imputación (considerando solamente registros completos, excluyendo los registros con datos faltantes) y luego de la aplicación de cada uno de los métodos de imputación (incluyendo ahora los archivos completos, es decir los registros con datos ahora imputados), para lo cual habrá de definirse una métrica específica.

La metodología desarrollada será general, pero la aplicación de la misma será particular, para cada archivo con datos faltantes donde sea necesario imputar los mismos. La aplicación de la metodología general a desarrollar será parte de un modelo de decisión que se aplicará ante casos concretos de archivos con datos faltantes (la metodología se desarrollará como una herramienta en el modelo se decisión que incluye la especificación del contexto en el que se aplicará la metodología desarrollada).

Los objetivos específicos son:

- Definir la métrica a utilizar para la validación de los diferentes métodos de imputación aplicados al conjunto de datos.
- Definir el procedimiento de selección de los métodos de imputación más adecuados para ser aplicados al conjunto de datos.
- Definir el orden de prioridad de las variables para la aplicación de los diferentes métodos de imputación de datos en el conjunto de datos.

3. Formación de Recursos Humanos

El equipo de trabajo está compuesto por un Doctor, dos Magister y dos Licenciados en Sistemas de Información con cursados de

Maestría Finalizadas, de los cuales uno está desarrollando su Tesis en la propuesta presentada.

Referencias

[1] D. L. la Red Martínez, J. C. Acosta, V. E. Uribe y A. R. Rambo, "Academic Performance: An Approach From Data Mining," *Journal of Systemics, Cybernetics and Informatics,* vol. 10, no. 1, pp. 66-72, 2012.

[2] D. L. La Red Martínez, M. Karanik, M. Giovannini, N. Pinto, "Academic Performance Profiles: A Descriptive Model Based on Data Mining," *European Scientific Journal,* vol. 11, no. 9, pp. 17-38, March 2015.

[3] D. L. La Red Martínez, G. Bobadilla Almada, "Estudio del rendimiento académico y detección temprana de perfiles de alumnos en la Facultad Politécnica de la Universidad Nacional del Este de Paraguay," *en Utilizando tecnologías en la educación para fortalecer la práctica docente en América Latina. Revisiones teóricas - Experiencias prácticas*, Bogotá, 2016.

[4] D. L. La Red Martínez, M. Karanik, M. Giovannini, R. Scappini, "Towards to a Predictive Model of Academic Performance Using Data Mining in the UTN-FRRe," *Journal of Systemics, Cybernetics and Informatics,* vol. 14, no. 2, pp. 36-41, 2016.

[5] D. L. La Red Martínez, M. E. Giovannini, M. E. Báez Molinas, J. I. Torre, N. Yaccuzzi, "Academic performance problems: A predictive data mining-based model," *Academia Journal of Educational Research,* vol. 5, no. 4, pp. 61-75. April 2017.

[6] U. Fayyad, G. Piatetsky-Shapiro & P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine,* vol. 17, no. 3, pp. 37-54, March 1996.

[7] J.I. Peláez, J. M. Doña, D. L. La Red Martínez, "Fuzzy Imputation Method For Database Systems," *in Handbook of Research on Fuzzy Information Processing in Database,* Hershey, PA: IGI Global (701 E. Chocolate Avenue, Hershey, Pennsylvania, 17033, USA), 2008.

[8] G. Madhu and T. V. Rajinikanth, "A novel index measure imputation algorithm for missing data values: A machine learning approach," *in 2012 IEEE International Conference on Computational Intelligence and Computing Research,* Coimbatore, 2012, pp. 1-7.

[9] J. Lengo, S. García & F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowledge and Information Systems,* vol. 32, issue 1, pp. 77-108, July 2012.

[10] A. Farhangfar, L. A. Kurgan and W. Pedrycz, "A Novel Framework for Imputation of Missing Values in Databases," *in IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* vol. 37, no. 5, pp. 692-709, Sept. 2007.

[11] M. Pattanodom, N. Iam-On & T. Boongoen, "Clustering data with the presence of missing values by ensemble approach," *in 2016 Second Asian Conference on Defence Technology (ACDT),* Chiang Mai, 2016, pp. 151-156.

[12] G. Rahman, Z. Islam, "A Decision Tree-based Missing Value Imputation Technique for Data Pre-processing," *in AusDM '11 Proceedings of the Ninth Australasian Data Mining Conference,* Ballarat, 2011, pp. 41-50.

[13] P. Schmitt, J. Mandel & M. Guedj, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics & Biostatistics,* vol. 6, issue 1, January 2015.

[14] Y. Liu & V. Gopalakrishnan, "An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data," *Data,* vol. 2, no. 1, p. 8, Jan. 2017.

[15] T. Aljuaid & S. Sasi, "Proper imputation techniques for missing values in data sets," *in 2016 International Conference on Data Science and Engineering (ICDSE),* Cochin, 2016, pp. 1-5

[16] B. Twala, M. Cartwright and M. Shepperd, "Comparison of various methods for handling incomplete data in software engineering databases," *in 2005 International Symposium on Empirical Software Engineering,* 2005, pp. 105-114.

[17] Y. Vergouwe, P. Royston, K. G.M. Moons, D. G. Altman, "Development and validation of a prediction model with missing predictor data: a practical approach," *Journal of Clinical Epidemiology,* vol. 63, issue 2, 2010, pp. 205-214.

[18] James L. PeughCraig K. Enders, "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement," *Review of Educational Research,* vol. 74, issue 4, 2004, pp. 525 – 556.

[19] G. Tutz & S. Ramzan, "Improved methods for the imputation of missing data by nearest neighbor methods," *Computational Statistics & Data Analysis,* vol. 90, 2015, pp. 84-99.

[20] A. Pantanowitz & T. Marwala, "Evaluating the Impact of Missing Data Imputation through the use of the Random Forest Algorithm," https://arxiv.org/abs/0812.2412, Fecha de Acceso 18 de Diciembre de 2017.

[21] K. Arima, N. Okada, Y. Tsuji and K. Kiguchi, "Evaluations of a multiple SOMs method for estimating missing values," *in 2014 IEEE/SICE International Symposium on System Integration,* Tokyo, 2014, pp. 796-801.

[22] D. B. Rubin, "Inference and Missing Data," *Biometrika,* vol. 63, issue 3, 1976, pp. 581-592.