Procesamiento de bases de datos escolares por medio de redes neuronales artificiales

García, Brenda Miranda; González Bárcenas, Víctor Manuel; Reyes Nava, Adriana; Alejo Eleuterio, Roberto; Rendón Lara, Eréndira

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Procesamiento de bases de datos escolares por medio de redes neuronales artificiales

School database Processing from the perspective of artificial neural networks

Brenda Miranda García
Instituto Tecnológico de Toluca, México
engineer.brend@gmail.com
http://orcid.org/0000-0001-6954-4581

Víctor Manuel González Bárcenas
Instituto Tecnológico de Toluca, México
sistemasvmanuel@gmail.com
http://orcid.org/0000-0002-9009-9686

Adriana Reyes Nava
Tecnológico de Estudios Superiores de Jocotitlán, México
adriananava0@gmail.com
http://orcid.org/0000-0002-4440-909X

Roberto Alejo Eleuterio
Instituto Tecnológico de Toluca, México
ralejoe@toluca.tecnm.mx
http://orcid.org/0000-0002-7580-3305

Eréndira Rendón Lara
Instituto Tecnológico de Toluca, México
erendonl@toluca.tecnm.mx
http://orcid.org/0000-0003-4581-6022

Recepción: 22 de octubre de 2019
Aprobación: 20 de mayo de 2020

Resumen

El estudio de bases de datos escolares es un área que ha sido poco estudiada y cuestionada desde el punto de vista de la minería de datos o de la inteligencia artificial. Actualmente, existen algunos trabajos que muestran su procesamiento mediante algoritmos de aprendizaje automático o "inteligentes"; sin embargo, no se detienen en analizar la pertinencia de procesar datos cualitativos como si fueran cuantitativos. En este artículo se estudia este problema con el uso de tres modelos de red neuronal. Los resultados evidencian la capacidad de estos modelos para clasificar con un porcentaje de acierto superior a 95% las tendencias en los estudiantes utilizando principalmente datos cualitativos.

Palabras clave: inteligencia artificial, redes neuronales artificiales, bases de datos escolares y datos cualitativos.

Abstract

The analysis of school mentoring databases is a poorly studied area and it is usually questioned from the point of view of data mining or artificial intelligence. Nowadays, there are some works about the processing of such a type of databases through machine learning algorithms, as well as the so called "smart algorithms". However, the relevance of analyzing and processing qualitative data as if they were quantitative remains still interesting. In this research, the problem of analyzing school mentoring databases by means of three artificial neural network models are thoroughly studied. Results shows the ability of these models to classify the correct trends in students' statistics using mainly qualitative data with a high degree of certainty (more than 95% of accuracy).

Keywords: Artificial intelligence, artificial neural network, mentoring database and qualitative data.

Introducción

Actualmente, en las Instituciones de Educación Superior (IES) de México se están generando miles de registros de información acerca de los estudiantes inscritos en los diferentes programas educativos que abarcan desde antecedentes académicos hasta familiares y socioeconómicos; estos últimos muchas veces son cualitativos. Asimismo, con el incremento de la matrícula de estudiantes se ha reducido la disponibilidad de recursos humanos para el análisis y estudio de esta información, lo cual impide su adecuada explotación y uso, así como una toma de decisiones basada en el conocimiento implícito en los datos de los estudiantes.

En los últimos años se han hecho esfuerzos por tratar de procesar y obtener conocimiento de los datos de los estudiantes por medio de tecnologías emergentes como la minería de datos, el aprendizaje automático (*Machine Learning*) y recientemente el profundo (*Deep Learning*). Estos estudios han sido dirigidos principalmente a investigar la deserción escolar (Vásquez, Peláez, y Ochoa, 2015; Jiménez y Tamiran, 2015). Eckert y Suénaga (2015) proponen la creación de un modelo basado en árboles de decisión para establecer la relación entre los factores que causan la deserción estudiantil. De manera semejante en Tan y Shao (2015) y Oviedo, Zambrano-Vega y Gómez (2019) se aborda este problema. Heredia, Amaya y Barrientos (2015) utilizaron los prototipos difusos mediante clúster, una técnica de clasificación no supervisada a través de la cual se realizó la extracción de variables descriptivas del rendimiento académico de los estudiantes. De forma similar, se desarrollaron los trabajos presentados en Yukselturk, Ozekez y Türel (2014) y Solís *et al*. (2018). Mendiola *et al*. (2015) describen diferentes métodos de clasificación de datos como *clustering*, reglas de asociación y árboles de decisión que usan información del rendimiento de los estudiantes para determinar por qué los alumnos abandonan los estudios. Solís *et al*. (2018) abordan el problema de alta dimensionalidad y desbalance de datos en la predicción de falla estudiantil. Ambos, dimensionalidad y desbalance, son retos importantes y de actualidad en el aprendizaje automático e inclusive en el profundo o *Deep Learning* (Leevy *et al*., 2018).

La deserción escolar también ha sido estudiada en áreas como las redes neuronales artificiales; por ejemplo, en los trabajos presentados en Tan y Shao (2014), López, Alejo y Velázquez (2015), Reyes *et al.* (2017), Miranda y Guzmán (2017), Naser *et al*. (2015) y Martin *et al*. (2018) se estudiaron datos históricos de estudiantes por medio de redes neuronales para clasificar posibles desertores. No obstante, en todos estos trabajos no se pone énfasis en el tipo de datos utilizados, es decir, si son cuantitativos o cualitativos.

Este artículo tiene como objetivo evaluar la efectividad de tres modelos de red neuronal artificial en la clasificación de alumnos de alto y bajo rendimiento, así como alumnos en riesgo de deserción. La base de datos utilizada proviene de investigaciones previas (López, Alejo y Velázquez *et al*., 2015; Reyes *et al*., 2017) desarrolladas en la carrera de Ingeniería en Sistemas Computacionales (ISC) del Tecnológico de Estudios Superiores de Jocotitlán (TESJo). Cabe destacar que esta investigación es relevante porque se centra en el procesamiento de datos cualitativos a través de mecanismo cuantitativos como las redes neuronales artificiales.

1. Trabajos relacionados

Hoy en día se han realizado diferentes investigaciones donde se trata de identificar cuáles son los factores que ocasionan que un alumno deserte por medio de métodos de inteligencia artificial como minería de datos y aprendizaje automático. Asimismo, se han realizado modelos basados en estas estrategias, donde se distinguen grupos con características similares que permiten la identificación y clasificación de alumnos en probable riesgo de deserción.

La detección de factores para tratar el problema de la deserción se puede hacer mediante diferentes algoritmos y usando información representativa de la base de datos como se presenta en Jiménez y Tamiran (2015), donde se realizó la detección de patrones a partir de datos socioeconómicos, académicos, disciplinares

e institucionales de los alumnos para obtener conocimiento que ayude en la toma de decisiones enfocadas en las políticas y estrategias relacionadas con los programas de retención estudiantil. Este estudio se realizó utilizando los algoritmos J48, Apriori y *K-means* y los resultados obtenidos determinaron un factor común para los estudiantes que desertan, el cual está relacionando con bajos promedios y materias perdidas en los primeros semestres de la carrera.

Algo similar es lo presentado por Eckert y Suénaga (2015), donde se analiza la información de estudiantes de la carrera de Ingeniería Informática para encontrar factores detonantes de deserción, se sigue la metodología KDD y se aplican técnicas de minería de datos como árboles de decisión (J48), redes bayesianas (*BayesNet*) y reglas OneR demostrando que existe una mayor tendencia para abandonar la carrera en alumnos de primer año en comparación con los más adelantados.

En Vásquez, Pelaez y Ochoa (2015) se utilizó una técnica de clasificación no supervisada llamada prototipos difusos mediante clúster a través de la cual se realizó la extracción de variables descriptivas del rendimiento académico de los estudiantes; la investigación fue realizada con base en la carga semestral de los alumnos para predecir posibles fallas del rendimiento académico en un punto de la carrera.

En los trabajos realizados en Yukselturk, Ozokez y Türel (2014), Heredia, Amaya y Barrientos (2015), Mendiola *et al*. (2015), Lakkaraju *et al*. (2015) y Tan y Shao (2015) se implementó un modelo predictivo mediante árboles de decisión principalmente. Al final de la construcción y ejecución del modelo propuesto se validaron los resultados comparando con la información de la base de datos original, la cual ya tenía identificados a los alumnos desertores y, de este modo, se pudo predecir quienes eran los estudiantes con probabilidad de desertar.

En la mayoría de los trabajos consultados los autores se centran en la identificación de características relevantes o en la predicción de estudiantes con probabilidades de desertar y no profundizan en la pertinencia de los datos utilizados. No obstante, algunos recientes (Oviedo, Zambrano-Vega y Gómez , 2019) analizan la posibilidad de usar modelos, como el clasificador bayesiano, para determinar la deserción en estudiantes tomando en cuenta datos socioeconómicos, es decir, considerando el tipo de dato empleado. Los resultados presentados muestran que no es posible determinar los alumnos que desertan usando ese tipo de algoritmos con la información que analizan.

2. Redes neuronales artificiales

Una Red Neuronal Artificial (RNA) es un modelo matemático que trata de emular a los sistemas neuronales biológicos del ser humano en el procesamiento de la información (Haykin, 2009). El modelo de RNA más común es el de propagación hacia adelante, el cual recibe información de su ambiente y la propaga a través de sus capas para tener como salida una predicción o una clasificación. Se caracteriza porque en ella no existen ciclos o conexiones hacia atrás. Los modelos más comunes de RNA de propagación hacia adelante son el perceptrón multicapa (*Multilayer Perceptron*-MLP), las redes de funciones de base radial (*Radial Base Function*-RBF) y las máquinas de vectores soporte (*Support Vector Machine*-SVM). Sin embargo, estas RNA se diferencian principalmente porque las RBF y SVM cuentan con una sola capa oculta que permite transformar el espacio de entrada a otro de diferente dimensión, a diferencia del MLP que puede tener varias capas.

Asimismo, el enfoque de entrenamiento que usan es distinto, en el caso del MLP y las RBF lo más común es usar la minimización del error en la separación de las clases o en los datos a los cuales es ajustado el modelo, mientras que en las SVM se maximiza la distancia entre las distintas clases. No obstante, tiene en común muchas características como ser clasificadores universales, realizar una transformación del espacio de entrada, son clasificadores lineales y mecanismos supervisados, pero sobre todo son modelos que trabajan en exclusivo con números reales (Haykin, 2009). Actualmente, las RNA, MLP, RBF y SVM son muy populares en áreas de aprendizaje automático, minería de datos y reconocimiento de patrones en tareas de clasificación, regresión y predicción (Duda, Hart y Stork, 2001).

3. Metodología

En esta sección se presentan las principales etapas y características experimentales más relevantes de la evaluación de la pertinencia del procesamiento de datos cualitativos a través de mecanismo cuantitativos como las redes neuronales artificiales, en particular de los modelos MLP, RBF y SVM; para ello, se definieron las siguientes etapas: *a*) adquisición de los datos, *b*) marco de trabajo a usar y *c*) mecanismos de evaluación del modelo neuronal.

Los datos usados para la clasificación de alumnos se obtuvieron de la aplicación de cuestionarios a estudiantes del TESJo de la carrera ISC de las generaciones 2010-2014 (López, Alejo y Velázquez, 2015) y son almacenados en una base de datos relacional, donde posteriormente se seleccionan aquellas tablas que contienen la información relevante sobre los alumnos.

El total de tablas que conforman la base de datos es de 60, de las cuales se obtienen 230 atributos y 1 161 muestras, de aquí se disminuye el número de tablas y atributos con los que se trabaja. La base de datos usa relaciones con claves primarias; de tal manera, del total de tablas, 17 de ellas contienen información utilizada dentro del cuestionario como menús desplegables (respuestas predefinidas) y las preguntas realizadas a los alumnos, razón por la cual no son relevantes para ser usadas dentro del análisis de los datos. Las respuestas a preguntas abiertas no se consideraron importantes para esta investigación porque se tendrían que procesar de manera distinta. Por ejemplo, en "me veo al finalizar la carrera como un profesionista más o menos exitoso" sería difícil asociar esta respuesta a un valor numérico. En el área de análisis de sentimientos con aprendizaje profundo, ya se están abordado temáticas de este tipo (Agarwal *et al*., 2020); sin embargo, el tratamiento de este tipo de información queda fuera del objetivo de este trabajo. Entonces, al final se obtienen 43 tablas en las que se encuentran los atributos de interés.

El formato original de la base de datos se encuentra en un archivo SQL, posteriormente se convierte a un archivo en formato CSV (archivo de texto separado por comas), donde se coloca la información de antecedentes académicos, sociales y familiares del alumno en una sola fila. El total de atributos que se obtienen de este archivo es de 127; sin embargo, para procesar la información, se eliminan 50 atributos como número de control, nombre de los estudiantes, entre otros, los cuales consideramos que no son de relevancia porque se quiere obtener conclusiones generales y no particulares. Finalmente, se obtuvieron 77 atributos a partir de los cuales se llevó a cabo el trabajo de minería de datos, 19 de ellos tienen valores numéricos (cuantitativos) y 58 tienen valores nominales (cualitativos) (tabla 1).

Asimismo, la base de datos final fue dividida en cuatro clases: *a*) alumnos desertores (clase 0), *b*) en riesgo de deserción (clase 1) 299 muestras, *c*) con alto rendimiento (clase 2) y *d*) en situación desconocida (clase 3). Este proceso se realizó con base en los resultados de investigaciones previas (López, Alejo y Velázquez, 2015; Reyes *et al*., 2017) sobre esta base de datos, y un especialista humano la llevó a cabo para corroborar la asignación de etiquetas a cada estudiante.

Las RNA son modelos cuantitativos, los cuales realizan para su funcionamiento operaciones en los datos de entrada como sumas, multiplicación y exponenciación e inclusive el uso de funciones trigonométricas (sección 2), es decir, operan sobre números reales. Por lo tanto, fue necesario hacer una transformación de los datos cualitativos a cuantitativos. Para esto, se asignó un valor numérico a los datos nominales, por ejemplo, en respuestas como "sí" y "no" se reemplazaron con 1 y 2, respectivamente.

En el cuadro 1 se muestra un ejemplo de la transformación realizada a los atributos cualitativos de la base de datos, la cual consintió en ordenar las respuestas: a la primera se le dio el valor de 1, a la segunda de 2 y así sucesivamente. Por ejemplo, el sexo es 1 para M (masculino) y 2 para F (femenino); para la fecha de nacimiento, sólo se dejó en año, estado civil es casado(a) --1, soltero(a) --2, divorciado(a) --3, unión libre --4, viudo(a) --5, separado(a) --6 y otro --7. El 7 se refiere a situaciones donde el estudiante no cree entrar en alguna de las categorías preestablecidas o no quiere dar esa información.

TABLA 1
Atributos de la base de datos de tutorías del TESJo

Cualitativos			Cuantitativos
Asesor	Tipo_Enferme	PqPromedio	Año
Sexo	Control_Medico	Tecestusino	Día
Mes	Medicamento	Tecestudio	Semestre
Edo_Civil	Nodependientes	Pq_trabaja	Duración_Est
Licenciatura	Hermanos_tiene	Fam_integrada	Promedio
Esc_Proced	Sosten_Estudios	Comoeslagente	Prom_ult_sem
Motivoaplaz	Trabajo	Relaciongente	Promgraltec
Motivos	ProblemEconom	Salud_genereal	Lugar_Ocupas
Cual_Lic	Transporte	Fumas_sino	Hermanos_estud
Motivo_Lic	Compu_Propia	Frec_fumas	Miembrconingre
Opcions_Lic	Internet	Tomas_sino	Tiempo_trab
Cualgustestud	Id_opcion	Frec_tomas	Dependientesdeti
ProyConclu	Fk_Periodo	Ofr_dorgasino	Numero_Hijos
Doncampolab	Id_Deporte	Probaste_drog	Codigo_Postal
Ayuda_Traba	Estado	Id_hobbie	Traducir
HermanosLic	Municipio	Id_Idiomas	Escribir
Rela_Padres	Nacionalidad	Parentesco	Leer
Rela_Herm	Porque_Carrera	Ocupación	Hablar
Rela_Casado	Porque_Instit	Id_vivi_con	Edad_padre
			Nivel_Estudios

Fuente: elaboración propia.

CUADRO 1
Ejemplos de la transformación de los atributos

Datos	Sexo	Fecha de nacimiento	Estado Civil	Escuela de procedencia	Duración de estudios (semestres)
Original	m	1981-07-14	f	3368	c
Transformado	1	1981	7	79	9
Original	F	1993-05-23	a	4168	b
Trasformado	2	1993	5	89	5

Fuente: elaboración propia.

Asimismo, por simplicidad, cuando se presentaron registros con datos faltantes se les asignó un valor no significativo a la base de datos, el cual es representado por 0; el *software* Weka también lo hace de esta manera, el cual asigna un valor por defecto de 0 a los datos faltantes. Este método es considerado como un método de imputación (Han, Kamber y Pei, 2012), donde se sustituye el valor faltante por una constante global.

Los atributos seleccionados almacenan únicamente respuestas a preguntas cerradas y la información de preguntas abiertas no se incluyó porque consideramos que quedan fuera del objetivo de esta investigación.

Por otro lado, para evaluar la efectividad de las RNA estudiadas en este artículo, se utilizó como medida de clasificación la bases de datos de estudiantes (la cuales contienen tanto valores cuantitativos como cualitativos) Múltiple área bajo la curva ROC (MAUC), la cual es una variante del área bajo la curva ROC (Receiver Operating Characteristic), pero para problemas de múltiples. MAUC puede ser definida como $MAUC = 2/|J|(|J| - 1)\sum_{i,j} AUC(j_i, j_k)$, donde $AUC(j_i, j_k)$ es el área bajo la curva de cada par de clases j_i, j_k (Fawcett, 2006).

Asimismo, como se observa en el cuadro 2, la base de datos estudiada tiene asociado un problema de desbalance de clases, es decir, una situación donde existen muchas más muestras de una clase que de otra, el cual ha sido reconocido como uno de los principales retos del aprendizaje automático, la minería de datos y el reconocimiento de patrones (Leevy *et al*., 2018).

CUADRO 2
Número de muestras en cada clase en la base de datos estudiada

Clase	Desertores (0)	En riesgo de deserción (1)	Alto rendimiento (2)	Situación desconocida (3)
Muestras	34	299	74	366

Fuente: elaboración propia.
Nota: alumnos desertores (clase 0), en riesgo de deserción (clase 1), de alto rendimiento (clase 2) y situación desconocida (clase 3).

Para abordar este problema se utilizaron las técnicas de muestreo ROS, RUS y SMOTE, las cuales han mostrado en numerosas ocasiones su efectividad para enfrentar este problema (López *et al*., 2013; Fernández *et al*., 2017). RUS se enfoca en el balance de las clases a través de la eliminación aleatoria de las muestras de las clases mayoritarias y ROS duplica muestras de las clases minoritarias con el propósito de balancear las clases (Patri, Batista, y Monard, 2009). SMOTE genera muestras artificiales de la clase minoritaria interpolando las muestras que están cercanas entre sí (Fernandez, García, y Herrera, 2018). Para cada muestra minoritaria, se encuentran los *k*-vecinos más cercanos y se generan muestras sintéticas con respecto a algunos o todos los vecinos más cercanos. ROS, RUS y SMOTE fueron aplicadas a la base de datos final, es decir, a la resultante de transformar los datos cualitativos en cuantitativos.

4. Resultados

En esta sección se presentan y discuten los resultados más relevantes de esta investigación. El cuadro 3 muestra los resultados de clasificar la base de datos de estudiantes (BDE), la cual fue preprocesada para adquirir un formato admisible por una RNA y para enfrentar el desbalance de clases (sección 3). Los resultados fueron obtenidos al aplicar la estrategia Validación-Cruzada (Refaeilzadeh, Tang y Liu, 2009) con cinco particiones, es decir, $BDE = BDE_1 \cup BDE_2 \cup BDE_3 \cup BDE_4 \cup BDE_5$ en donde cada partición BDE_i fue usada para la evaluación del modelo y el resto de las particiones ($BDE_{j \neq i}$, donde $i, j = 1, 2, ..., 5$) para el entrenamiento.

La tabla 4 presenta el $AUC_{p, q,}$ donde q representa la unión de todas las clases excepto la clase p, i. e., es el área bajo la curva ROC de la clase p con respecto al resto de las clases. MAUC se calcula como se indicó en sección anterior y representa el promedio del AUC para todas las clases.

Los resultados presentados en el cuadro 3 indican que los modelos de RNA estudiados en este artículo son eficientes para clasificar BDE con valores cualitativos. Se observa en la primera columna (ORIGINAL), que los valores de AUC e inclusive de MAUC son superiores a 80% de acierto para las cuatro clases establecidas en la BDE, con los modelos MLP y RBF. Sin embargo, para la clase 2 con SVM se obtiene un valor de 57.2 % de AUC o de acierto, pero cuando se aplican las técnicas para tratar el desbalance de clases este valor es incrementado en al menos 26%; en otras palabras, la baja efectividad mostrada por el modelo SVM para la clase 2 se debía al

problema del desbalance de clases y no a la naturaleza de los datos.

Asimismo, se observa en los resultados que los métodos RUS, ROS y SMOTE son apropiados para tratar el desbalance de clases en este tipo de bases de datos; no obstante, al igual que en otros trabajos (Alejo *et al*., 2017), el método RUS es el que produce peores resultados debido a la eliminación de muestras que realiza este procedimiento. ROS y SMOTE presentan resultados similares, aunque se advierte una tendencia: SMOTE genera mejores resultados que ROS, lo cual ya ha sido reportado en la literatura especializada (Fernandez, García y Herrera, 2018).

Por otra parte, es interesante notar que la RNA que obtiene mejores resultados es MLP con una capa oculta, lo cual puede deberse a que el MLP realiza una transformación del espacio de acuerdo con los atributos de entrada, mientras que las SVM y RBF la transformación es por prototipo o muestra utilizando una función núcleo, que generalmente es una función gaussiana, lo cual trae implícito un comportamiento normal de los datos de entrada.

CUADRO 3
Valores de AUC para pares de clases y promedio general

Clase	ORIGINAL			SMOTE			ROS			RUS		
p	MLP	SVM	RBF	MLP	SVM	RBF	MLP	SVM	RBF	MLP	SVM	RBF
0	0.973	0.954	0.828	0.998	0.984	0.985	0.993	0.993	0.972	0.895	0.866	0.741
1	0.929	0.884	0.929	0.981	0.912	0.886	0.971	0.939	0.937	0.877	0.88	0.708
2	0.876	0.572	0.909	0.99	0.958	0.942	0.984	0.949	0.974	0.858	0.835	0.863
3	1	1	0.997	1	1	1	1	1	1	1	1	1
MAUC	0.96	0.92	0.955	0.992	0.974	0.95	0.987	0.97	0.971	0.881	0.878	0.83

Análisis prospectivo

Los resultados presentados en la sección 4 evidencian la efectividad de los modelos de RNA, SVM, RBF y MLP para clasificar bases de datos con valores cualitativos (a pesar de que la conversión de cualitativo a numérico fue muy simple e inclusive burda). En la sección 2 se discutió sobre la naturaleza cuantitativa de los modelos de RNA estudiados en este trabajo y en la sección 4 de cómo los datos cualitativos de la base de datos fueron cuantificados.

La discusión aquí es la pertinencia de usar modelos cuantitativos para tratar datos cualitativos. En la base de datos de este artículo, la mayor parte de los atributos son cualitativos (tabla 1). No obstante, los resultados muestran que el uso de modelos de RNA son apropiados para la predicción de estudiantes desertores (clase 0), en riesgo de deserción (clase 1), con alto rendimiento (clase 2) y en situación desconocida (clase 3).

Por otro lado, se debe cuestionar el funcionamiento de la RNA cuando los valores usados representan situaciones como si los estudiantes fuman o no, si tienen una buena relación con sus padres, su escuela de origen, entre otros, porque éstos son multiplicados, divididos e incluso son parámetros de funciones exponenciales. Las preguntas son "¿por qué funcionan estos modelos?", "¿es correcto lo que se está haciendo?", "¿debería hacerse de otro modo?".

En este sentido, consideramos que los modelos funcionan porque encuentran correlaciones entre los datos debido a frecuencia de aparición de patrones en los datos. En otras palabras, encuentran tendencias o comportamientos en los objetos de estudio, independientemente de su naturaleza; por ejemplo, puede haber características en común entre los desertores como la relación que tienen con sus padres, las cuales pueden ser ubicadas por la RNA. No obstante, por la naturaleza de *caja negra* (Benítez *et al*., 1997) de las RNA, este conocimiento no se puede observar explícitamente en la RNA. Por otro lado, creemos que si la RNA clasifica bien, entonces es correcto usarlas para este tipo de actividades; sin embargo, es necesario seguir profundizando y desentrañando los secretos de las RNA. Asimismo, pensamos que tratar los datos de esta forma es una tendencia que seguirá creciendo y se consolidará en ambientes como los de Big Data y la Inteligencia Artificial contemporánea debido a que cada día se generan más y más datos a una gran velocidad y diversidad, y procesarlos de manera tradicional es una tarea prácticamente imposible.

Conclusiones

Este artículo se encargó de estudiar la viabilidad de clasificar bases de datos con valores cualitativos por medio de modelos puramente cuantitativos, i. e., las redes neuronales artificiales. La base de datos utilizada corresponde a un concentrado de información proveniente del sistema de tutorías del Tecnológico de Estudios Superiores de Jocotitlán. Las redes neuronales utilizadas fueron los modelos máquinas de vectores soporte, las redes RBF y el perceptrón multicapa.

Los resultados presentados muestran la efectividad de los modelos neuronales para clasificar bases de datos, en las cuales la mayoría de la información es cualitativa; por ejemplo, la relación de los estudiantes con sus padres, pasatiempos, si fuman, su escuela de procedencia, sexo, entre otros. Asimismo, se observó que el desbalance de clases afecta el rendimiento del clasificador, inclusive en este contexto, y que los métodos de muestreo tradicionales ROS y SMOTE son efectivos para enfrentar ese problema.

Finalmente, se dan respuestas a preguntas abiertas acerca de la validez de estos modelos matemáticos para tratar datos cualitativos, más allá de su efectividad o precisión en la clasificación. La principal conclusión a la que se llegó en este trabajo es la utilidad e importancia de estos modelos matemáticos para tratar valores cualitativos.

Es indudable que se requiere ampliar esta investigación no sólo por su importancia, sino por su vinculación con mecanismos automatizados para la toma de decisiones en contextos con un alto componente cualitativo. Asimismo, se buscará explorar conceptos como variables indicadoras en la transformación de variables cualitativas a cuantitativas, i. e., desde el punto de vista de la econometría.

Agradecimientos

Se agradecen los comentarios de los árbitros de la revista que mejoraron sustancialmente el contenido del artículo.

Referencias

Alejo, R., Monroy, J., Ambriz, J. C., & Pacheco-Sánchez, J. H. (2017). An improved dynamic sampling back-propagation algorithm based on mean square error to face the multiclass imbalance problem. *Neural Computing and Applications*, *28*(10), 2843-2857.

Agarwal, B., Nayak, R., Mittal, N., & Patnaik, S. (2020). *Deep learning-based approaches for sentiment analysis*. Springer Singapore.

Benítez, J. M., Castro, J. L., & Requena, I. (1997). Are artificial neural networks black boxes? *IEEE transactions on neural networks*, *8*(5), 1156-1164.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern Classification and Scene Analysis*. New York.

Eckert, K. B. y Suénaga, R. (2015). Análisis de deserción permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos. *Formación universitaria*, *8*(5), 3-12.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861-874.

Fernández, A., del Río, S., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced big data classification: Outcomes and challenges. *Complex & Intelligent Systems*, *3*(2), 105-120.

Fernandez, A., García, S., & Herrera, F. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*(1), 863-905.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques*. Morgan Kaufmann Publishers.

Haykin, S. (2009). *Neural networks and learning machines.* Hamilton: McMaster University.

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student dropout predictive model using data mining techniques. *IEEE Latin America Transactions*, *9*(13), 3127-3134.

Jiménez, A. J. y Tamiran, R. S. (2015). Caracterización de la deserción estudiantil en educación superior con minería de datos. *Revista Tecnológica ESPOL*, *28*(5), 447-463.

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, L. K. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1909-1918.

Leevy, L. J., Khoshgoftaar, M. T., Bauder, A. R., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*(1), 1-42.

López, G. E., Alejo, R., & Velázquez, J. (2015). Loyalty index of students, a view from the Tesjo stage with data mining. *ECORFAN Journal*, *2*(2), 133-139.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113-141.

Mendiola, J. L. A., Valdovinos, R. M., Antonio, J. A. Alejo, R. y Marcial, R. (2015). "Análisis de deserción escolar con minería de datos," *Research in Computer Science*, *93*, 71-82.

Miranda, A. M. y Guzmán, J. (2017). Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. *Formación Universitaria*, 61-68.

Naser, A. S., Zaqout, I., Ghosh, A. M., Atallah, R. R., & Alajrami, E. (2015). Predicting student performance using artificial neural network: In the faculty of engineering and information technology. *International Journal of Hybrid Information Technology*, 221-228.

Oviedo, B., Zambrano-Vega, C., & Gómez, J. (2019). Clasificador bayesiano simple aplicado al aprendizaje. *RISTI E18*, 74-85.

Patri, R. C., Batista, G. E., & Monard, M. C. (2009). Data mining with imbalanced class distributions: Concepts and methods. *Proceedings of the 4th Indian International Conference on Artificial Intelligence* (359-376). IICAI: Tumkur.

Refaeilzadeh, P., Tang L., & Liu H. (2009) Cross-Validation. In L. Liu & M. T. Özsu (Eds.) *Encyclopedia of Database Systems*. Boston: Springer.

Reyes, N. A., Flores, F. A., Alejo, R. y Rendón, L. E. (2017). Minería de datos aplicada para la identificación de factores de riesgo en alumnos. *Research in Computing Science*, 177-189.

Solís, M., Moreira-Mora, T. E., González Laisa, R., Fernández-Martín, T., & Hernández-Jiménez, M. T. (2018). Perspectives to predict dropout in university students with machine learning. *In Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence* (*IWOBI*), 1-6.

Tan, M., & Shao, P. (2014). Predicting dropout from online education based on neural networks. *The Open Cybernetics and Systemics Journal*, 263-627.

Tan, M., & Shao, P. (2015). Prediction of student dropout in E-Learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning* (*iJET*), 11-17.

Vásquez, A., Peláez, E. y Ochoa, X. (2015). Predictor basado en prototipos difusos y clasificación no supervisada. *Revista Latinoamericana de Ingeniería de Software*, 135-140.

Yukselturk, E., Ozekez, S., & Türel, K. Y. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-Learning*, 118-133.