ISSN 1870-4069

Ensamble de clasificadores para determinar el perfil académico del estudiante usando árboles de decisión y redes neuronales

Maricela Quintana López, José Martín Flores Albino, Saúl Lazcano Salas, Víctor Manuel Landassuri Moreno

Centro Universitario UAEM, Valle de México, México

{mquintanal, jmfloresa, slazcanos, vmlandassurim}@uaemex.mx

Resumen. En este artículo, se propone el uso de un ensamble de clasificadores para determinar el perfil académico del estudiante, basado en su promedio general y en datos relacionados a los factores educativos: actividades de estudio, formas de aprendizaje y hábitos de estudio. Los datos usados se obtuvieron del cuestionario socioeconómico aplicado a los estudiantes del Centro Universitario UAEM Valle de México, asignándole la clase correspondiente de acuerdo con su promedio general. Las clases se definieron como excelente, bueno y regular. Para cada grupo de factores, se utilizó el algoritmo C4.5 para generar el clasificador correspondiente. El ensamble de clasificadores fue entonces diseñado utilizando una red neuronal artificial. La red neuronal recibe como entrada la clasificación asignada por los tres clasificadores y es entrenada para asignar la clase correcta usando un subconjunto de los datos. Se observa en los resultados que el ensamble propuesto tiene mejor desempeño comparado con los clasificadores independientes.

Palabras clave: Ensamble de clasificadores, árboles de decisión, redes neuronales artificiales.

Ensemble of Classifiers to Determine Student Academic Profile Using Decision Trees and Neural Networks

Abstract. In this paper, an ensemble of classifiers is proposed to determine student academic profile using decision trees and neural networks, based on his grade point average and data related to educative factors: study activities, learning forms and study habits. The data came from a socioeconomic questionnaire applied to the students of the University Center UAEM Valley of Mexico, and the corresponding class was assigned based on his grade point average. The classes were defined as: excellent, good, and regular. For each group of factors, the C4.5 algorithm was applied to build the corresponding

classifier. The ensemble of classifiers was designed using a neural network which receives the classes assigned by the three classifiers as input and it was trained to assign the corresponding class using a data subset. It is observed that the performance of the ensemble is better than the performance obtained for each independent classifier.

Keywords: Ensemble of classifiers, decision trees, artificial neural network.

1. Introducción

Una de las preocupaciones en las universidades es la deserción y el bajo desempeño académico de los estudiantes, es sabido que en esto influyen muchos factores, entre estos los sociales, económicos, educativos, académicos e institucionales, por citar algunos. En este trabajo se busca generar un ensamble de clasificadores para determinar el perfil académico de un estudiante, con base en su promedio general y usando como discriminantes los factores educativos como son: actividades de estudio, formas de aprendizaje y hábitos de estudio. La base de datos consiste en los registros de los estudiantes del Centro Universitario UAEM Valle de México que ingresaron en los años 2008, 2009 y 2010 en las diferentes carreras que se imparten.

Una estrategia popular es probar diferentes algoritmos para generar los modelos, evaluarlos y elegir el que proporcione los mejores resultados, esto es el que tenga el menor error en la predicción de la clase de una instancia desconocida. En contraste, un método de ensamble construye o utiliza varios clasificadores y los combina [1]. Usualmente se utilizan clasificadores construidos a partir de algoritmos diferentes sobre los mismos datos. Sin embargo, debido a que al hacerlo de esta manera, en este trabajo no llevó a buenos resultados, se optó por dividir los datos por grupos de factores y se aplicó el mismo algoritmo C4.5 construyéndose 3 clasificadores y generando el ensamble con ellos.

El documento contiene la siguiente estructura, en la sección 2, se presenta la metodología empleada para realizar el presente trabajo, mientras que en la sección 3 se muestra de manera muy general los algoritmos de minería de datos y redes neuronales empleados, así como el método de ensamble utilizado. En la sección 4, se presentan los experimentos y resultados obtenidos; finalmente en la sección 5, se muestran las conclusiones y trabajo futuro.

2. Metodología

Para conducir la investigación se realizaron etapas en las que la metodología utilizada fue la del proceso de extracción del conocimiento, conocido como KDD (Knowledge Discovery in Databases) [2], ver Figura 1 con esto:

a. Se creó y probó un clasificador considerando todos los factores educativos.

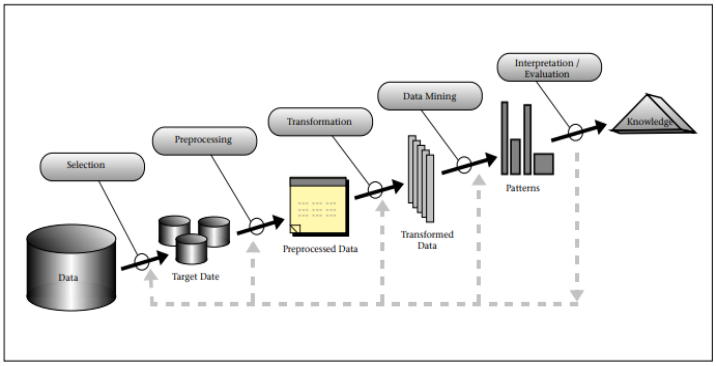

Fig. 1. Etapas del Proceso de Extracción del Conocimiento [2]

Tabla 1. Actividades de estudio.

Me reúno con mis compañeros para preparar un examen
Me reúno con mis compañeros para elaborar una tarea o un trabajo en grupo
Al iniciar, identifico lo que necesito estudiar y elaboro un plan de trabajo
Reviso qué es lo que recuerdo de lo que estudié
Identifico los conceptos que aún no he comprendido
Cuando no entiendo algo busco información esclarecedora
Estudio principalmente con monografías
Estudio principalmente con mis apuntes de clase
Estudio principalmente con el libro de texto de la asignatura
Estudio principalmente con los apuntes de mis compañeros
Uso enciclopedias, diccionarios o atlas
Uso computadora o Internet para estudiar, hacer tarea o resolver un examen

b. Se generaron clasificadores con bloques de factores educativos, es decir, actividades de estudio, formas de aprendizaje y hábitos de estudio y se probaron los modelos.
c. Se generó el ensamble utilizando redes neuronales artificiales.
d. Se analizaron y compararon los resultados.

2.1. Selección de los datos

Los datos fueron obtenidos del estudio socioeconómico perteneciente a los estudiantes de las generaciones 2008-2013, 2009-2014 y 2010-2015 de todas las carreras del Centro Universitario UAEM Valle de México. Los datos seleccionados corresponden a la sección de factores educativos, que se aplica a los estudiantes, específicamente a las actividades de estudio, ver tabla 1, formas de aprendizaje, ver tabla 2 y a los hábitos de estudio, ver tabla 3. Estos datos son relacionados con los datos de control escolar, específicamente con el promedio general del estudiante.

La escala de respuesta a las preguntas se divide en: *Nunca*, *Normal*, *Siempre*, excepto en aquellas donde se solicita un número de horas o de libros, que son las primeras 3 preguntas de los hábitos de estudio.

Tabla 2. Formas de aprendizaje.

Aprendo más cuando trabajo con otros compañeros
Es de gran ayuda que todos aporten ideas cuando trabajo en grupo
Estudio para asegurar económicamente mi futuro
Estudio para obtener un buen trabajo
Estudio para aprender más
Estudio para vivir mejor
Confío en que puedo entender lo que estudio inclusive los textos más difíciles
Confío en que puedo realizar un excelente trabajo en mis tareas y en exámenes
Tengo seguridad en que domino las habilidades que me enseñaron
Aprendo rápidamente en la mayoría de las asignaturas
Soy competente en la mayoría de las asignaturas
Resuelvo bien los exámenes en la mayoría de las asignaturas
Me gusta trabajar con otros compañeros
Solamente leo cuando tengo la obligación de hacerlo
La lectura es uno de mis pasatiempos favoritos
Me gusta comentar los libros con otras personas
Me cuesta trabajo terminar de leer un libro
Me gusta que me regalen libros
La lectura me parece una pérdida de tiempo
Disfruto el visitar librerías o bibliotecas
Solamente leo para obtener la información que necesito
Me cuesta trabajo sentarme a leer por mucho tiempo

Tabla 3. Hábitos de estudio.

¿Horas a la semana que estudia o hace tarea fuera del horario escolar?
¿Horas a la semana se dedica a leer sobre lo que le gusta o interesa?
Indique cuántos libros completos ha leído en los últimos 12 meses sin tomar en cuenta sus libros de texto
Libros de literatura (novela, teatro, poesía)
Libros de otros temas (ciencia, tecnología, economía, etc.)
Revistas
Periódicos
Historietas
Páginas de Internet

2.2. Preprocesamiento

En el archivo original la cantidad de instancias era de 3600, sin embargo, debido a que algunos estudiantes no contestaban una gran cantidad de preguntas, las instancias con información incompleta fueron eliminadas, quedando únicamente 1021 instancias.

Por otro lado, dentro de los datos, existen atributos que actúan únicamente como identificadores de las instancias, pero que no se utilizan para generar los clasificadores como el nombre del alumno, la edad, y el sexo.

2.3. Transformación

El perfil académico del estudiante se dividió en 3 clases, de acuerdo con el promedio general obtenido desde su ingreso y hasta el último semestre cursado. La tabla 4 presenta esta información, así como la cantidad de estudiantes que tienen el perfil.

Tabla 4. Distribución de alumnos por perfil académico.

Perfil Académico	Promedio General	Alumnos
Regular	[0.0,7.3]	168
Bueno	[7.4,8.7]	691
Excelente	[8.8,10]	162
	Total	**1021**

Tabla 5. Proporción de datos por perfil para los conjuntos de entrenamiento y prueba.

Conjunto	Perfil Académico			Total
	Regular	Bueno	Excelente	
Entrenamiento	136	553	129	**818**
Prueba	32	138	33	**203**
Total	**168**	**691**	**162**	**1021**

El conjunto de datos seleccionados, 1021 instancias o ejemplares, se dividió en dos conjuntos: el de entrenamiento con 818 instancias, y el de prueba con 203 instancias; con el fin de que todas las clases estuvieran representadas, se tomó aproximadamente el 20% de cada una para el conjunto de prueba y el resto de las instancias en el conjunto de entrenamiento (≈80%). La división de los datos se presenta en la tabla 5.

3. Minería de datos, redes neuronales y ensambles

A continuación se describe, de manera sucinta, lo referente a la minería de datos, y en especial, al algoritmo empleado en este trabajo; también se presentan los conocimientos básicos acerca de las redes neuronales y los ensambles.

3.1. Minería de datos

La minería de datos aparece como la tercera etapa del proceso de extracción del conocimiento, con ello se pueden realizar diferentes tipos de tareas que tiene como metas principales la predicción y la descripción [2]. De acuerdo con [3], la minería de datos se define como el proceso de encontrar información novedosa y comprensible a partir de los datos. En este trabajo, la información que se busca es la que indica que un alumno es excelente, regular o malo, y poder emplear, en un futuro, esta información tanto para clasificar a nuevos alumnos, como para entender qué actividades, hábitos y formas de aprendizaje definen a las diferentes clases, y con esta información crear

estrategias que permitan fomentarlas. El algoritmo de clasificación empleado en este trabajo es el C4.5., el cual es una mejora del algoritmo ID3 desarrollado por Ross Quinlan, permite trabajar con atributos numéricos y considera la información de la partición y el radio de ganancia, entre otras mejoras, para elegir los atributos ganadores; la salida es un árbol de decisión; Este algoritmo está implementado en el software libre WEKA de la Universidad de Waikato bajo el nombre de J48 [4] y es el que se utiliza en este trabajo.

3.2. Redes neuronales

Las redes neuronales artificiales son estructuras de cálculo que se basan en unidades relativamente simples de procesamiento, llamadas neuronas. Al interconectar las neuronas artificiales se multiplica su capacidad para representar conocimiento, entendido este último como una relación funcional compleja entre datos de entrada y datos de salida. La adquisición del conocimiento en una red neuronal depende del modelo de aprendizaje utilizado. Para problemas de clasificación se hace uso de un conjunto de aprendizaje que consiste en datos de entrada con su descriptor de clase conocido, a través de este conjunto se ajustan los parámetros de la red para que señalen la clase a la que se sabe que pertenecen. El proceso de actualización de los parámetros de la red neuronal es lo que se considera el entrenamiento. El entrenamiento hace uso de una medida del error de clasificación de la red (índice de desempeño) y por medio de una medida de la sensibilidad de los parámetros de la red al error, se realiza su adaptación, este suele basarse en el algoritmo llamado *backpropagation*.

Como se puede ver en [5], se presenta un panorama del uso de las redes neuronales en el análisis de grandes bases de datos. Las redes neuronales pueden construir modelos estadísticos no lineales a través los datos disponibles. El modelo va ajustándose para cumplir con el nivel del índice de desempeño establecido. Las tareas de clasificación consisten en asociar un grupo de patrones de entrada a la categoría que pertenecen y que los identifica o describe.

3.3. Métodos de ensamble

Los métodos de ensamble o combinación de modelos, surgen con el propósito de mejorar la precisión de las predicciones. Un ensamble contiene un número de aprendices (modelos base) que, cuando son del mismo tipo son llamados homogéneos y si son de diferentes, heterogéneos. La característica es que estos aprendices no tienen un buen desempeño. El ensamble se genera utilizando otro algoritmo que combina los aprendices, ejemplo de éstos son el voto mayoritario, la tabla de decisión y las redes neuronales [1].

La precisión obtenida por el ensamble o combinación de modelos, generalmente supera la precisión de cada componente. En [6] se presenta el esquema de combinación de modelos que tienen árboles de decisión como base, ver Figura 2. Se observa que los datos son dados a los árboles de decisión y cada uno entrega su clasificación, se realiza entonces la combinación y se entrega la predicción combinada.

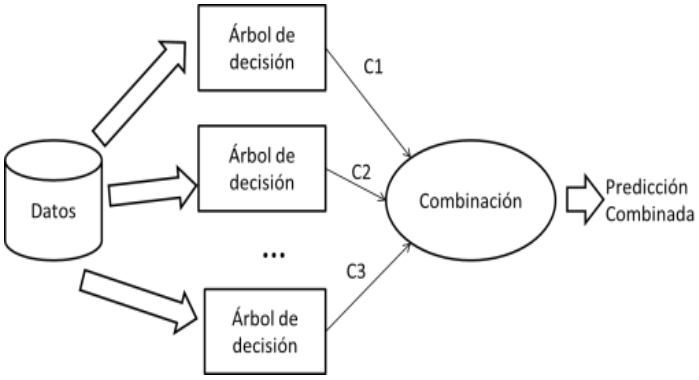

Fig. 2. Combinación de modelos usando árboles de decisión como base [6]

Los ensambles de clasificadores se han utilizado en ámbitos muy diversos, por ejemplo en [7], analizan diversos clasificadores detectores de malware para sistemas Android que basan su análisis en el uso de los recursos de hardware de los diversos programas en ejecución para detectar patrones sospechosos que sean característicos de programas malware. Realizan un ensamble final de dichos clasificadores, logrando que la detección exitosa de malware se incremente de manera significativa, comparado contra los clasificadores individuales.

Otro ejemplo de uso de ensambles de clasificadores es el trabajo de [8], en el cual analizan mensajes en la red social Twitter para clasificar los mismos en cuatro categorías. Los ensambles que proponen logran una clasificación de los mensajes con un grado de éxito más elevado comparado con otros clasificadores usados en esta tarea. Además, los ensambles que proponen tienen la característica de trabajar muy bien en situaciones donde el conjunto de entrenamiento no es lo suficientemente grande.

El trabajo presentado en [9], se enfoca en la detección de leucemia analizando imágenes de las muestras sanguíneas; dichas imágenes las pasan a través de una red neuronal convolutiva para la extracción de características, reducen el número de las mismas y seleccionan las más significativas y finalmente, las analizan a través de un ensamble de 3 clasificadores *(Support vector machine, multilayer perceptron, random forest*) logrando una detección exitosa prácticamente del 100% de los casos analizados, destacando la reducción del tiempo de análisis comparado con técnicas de detección de leucemia tradicionales.

En este trabajo se utiliza un ensamble de clasificadores para mejorar el desempeño de los clasificadores individuales; el ensamble se realizó utilizando Redes Neuronales.

4. Experimentos y resultados

A continuación se presentan los experimentos realizados y los resultados obtenidos, iniciando con la generación del clasificador único que utiliza todos los datos, prosiguiendo con los clasificadores individuales y terminando con el ensamble.

4.1. Generación y prueba de un clasificador considerando todos los factores educativos

El primer clasificador se generó utilizando el algoritmo C4.5. El modelo se generó a partir del conjunto de entrenamiento (818 instancias). Los resultados son presentados en la tabla 6, en términos de porcentajes de acierto y error en la clasificación de las instancias, de igual manera se presenta la matriz de confusión.

Tabla 6. Resultados del clasificador generado con el algoritmo C4.5 (J48 WEKA).

Clases	**Matriz de Confusión**			**Acierto**	**Error**
	A	**B**	**C**	89.12%	10.88%
a = Bueno	**538**	9	6	818 instancias	
b = Regular	29	**105**	2	729 clasificadas correctamente	
c = Excelente	35	8	**86**	89 clasificadas incorrectamente	

Tabla 7. Resultados del clasificador en el conjunto de prueba.

Clases	**Matriz de Confusión**			**Acierto**	**Error**
	A	**B**	**C**	58.62%	41.38%
a = Bueno	**102**	17	19	203 instancias	
b = Regular	19	**11**	3	119 clasificadas correctamente	
c = Excelente	24	2	**6**	84 clasificadas incorrectamente	

Al utilizar el modelo generado para clasificar las instancias en el conjunto de prueba, se puede observar que baja la eficiencia del clasificador de 89.12% a 58.62%. Los resultados del conjunto de prueba son presentados en la tabla 7.

4.2. Generación y prueba de clasificadores usando los factores educativos por separado: actividades de estudio, formas de aprendizaje y hábitos de estudio

Los datos fueron separados en actividades de estudio, formas de aprendizaje y hábitos de estudio. Con cada grupo se generó un clasificador usando el algoritmo C4.5 sobre el conjunto de entrenamiento y se calculó su desempeño con el conjunto de prueba. En la tabla 8 se presentan los resultados, es posible notar que, si bien los clasificadores mostraron un buen desempeño con el conjunto de entrenamiento, al probarlos la eficiencia al clasificar descendía.

Tabla 8. Resumen del desempeño de los clasificadores.

Factor Educativo	Entrenamiento		Prueba	
	Acierto	**Error**	**Acierto**	**Error**
Actividad de estudio	71.64	28.36	65.52	34.48
Formas de aprendizaje	75.18	24.82	64.53	35.47
Hábitos de estudio	78.24	21.76	62.07	37.93
Todos los factores	89.12	10.88	58.62	41.38

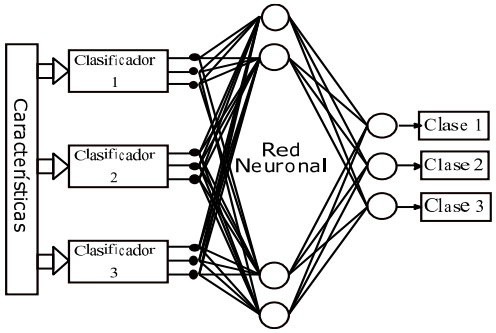

Fig. 3. Arquitectura de la Red Neuronal: 9 entradas y bias, 10 neuronas con función de activación *Logsig* y 3 neuronas de salida con función de activación *Logsig.*

A pesar de que en todos los factores educativos, los clasificadores individuales bajan el porcentaje de aciertos al utilizarse en el conjunto de prueba, estos tienen mejor desempeño que el clasificador único que contempla a todos los atributos (último renglón de la tabla 8).

4.3. Ensamble de clasificadores usando una red neuronal artificial

Arquitectura de la red neuronal (RN). La salida de los clasificadores se representa por un vector 3x1. P= *[Clase 1, Clase 2, Clase 3]*^{T}, de manera que la RN tiene nueve entradas. Las nueve entradas pasan a la primera capa de 10 neuronas, con de una entrada independiente de compensación (bias). La función de activación es de tipo logsig, (1). La capa final tiene tres neuronas con bias y función de activación logsig. El objetivo es que la salida represente la clase por medio del vector S= *[Clase1, Clase 2, Clase3]*^{T}, ver Figura 3:

$$Logsig(x) = \frac{1}{1-e^{-x}}. \quad (1)$$

Algoritmo de entrenamiento. Se utiliza el algoritmo de entrenamiento "El Gradiente Escalado Conjugado" [10]. El índice de desempeño es:

$$H_{y'}(y) = \sum_i y_i' \log(y_i), \quad (2)$$

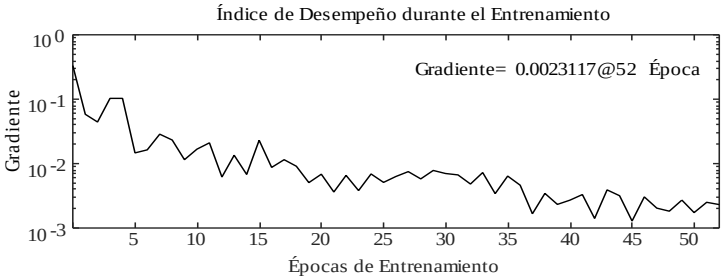

Fig. 4. Comportamiento del Índice de desempeño.

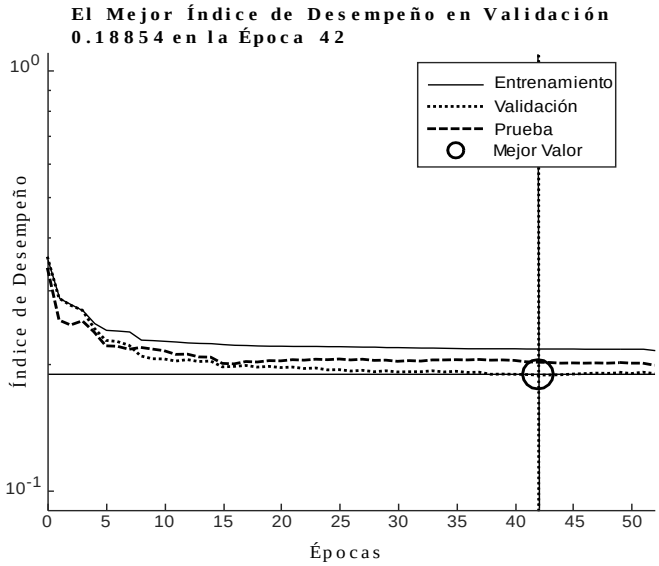

Fig. 5. Comportamiento del índice de desempeño con respecto a las fases de Entrenamiento, Validación y Prueba de la Red Neuronal.

donde:

y_i'es la probabilidad de la clase i que se espera,

y_i'es la probabilidad estimada.

Resultados experimentales. Al realizar el entrenamiento durante 52 épocas se obtuvieron los siguientes resultados. En la figura 4, se presenta el índice de desempeño de cada época de entrenamiento. El cálculo de este índice es a través de (2), que en la última época tiene un valor de 0.00123117.

En la Figura 5, se muestran tres series de valores de índice de desempeño para el grupo de datos tomados para entrenamiento, validación y prueba. Destaca que en la época 42 se obtuvo el mínimo del índice de desempeño igual a *0.18854* para el conjunto

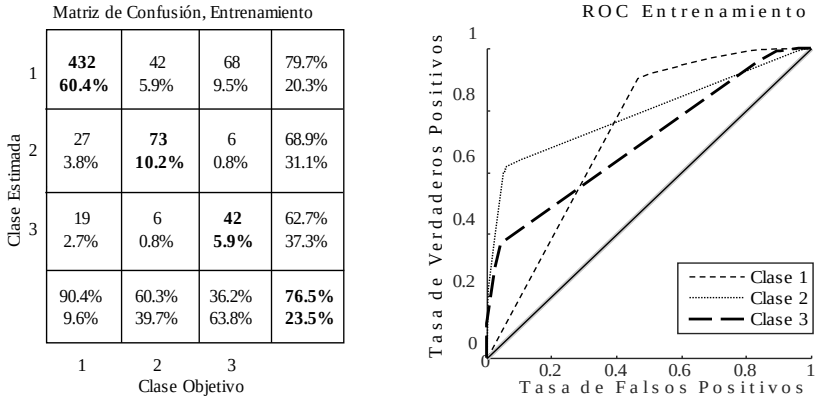

Fig. 6. (a) Matriz de confusión en la última etapa de entrenamiento. Porcentaje de predicción correcta del 76.5% para las tres clases. (b) Gráfico de "Receiver operating characteristic" Observar como en todos los resultados las curvas de sensibilidad están en la parte superior de la línea de no discriminación (diagonal).

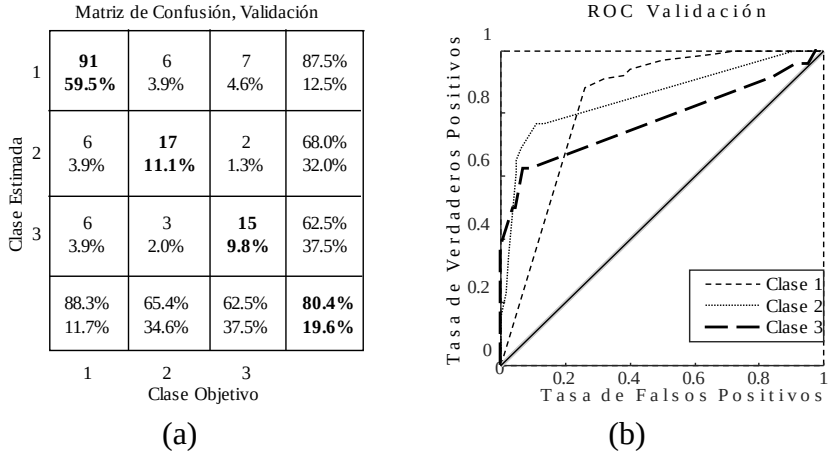

Fig. 7. (a) Matriz de confusión en la Validación. Porcentaje de predicción correcta del 80.4% para las tres clases. (b) Gráfico de "Receiver operating characteristic".

de datos de validación, esta información es tomada como medio de paro del algoritmo de entrenamiento.

Para evaluar la calidad del clasificador en las Figuras 6, 7 y 8 están las matrices de confusión y las gráficas ROC "Receiver operating characteristic" que permiten valorar la calidad de la estimación de clase. Durante el entrenamiento en la Figura 6, se observa un porcentaje de clasificación correcta para la clase 1 de 79.7%, para la clase 2 de 68.9% y de la clase 3 del 62.7%, dando un porcentaje promedio del 76.5%.

Al observar estos valores para los conjuntos de validación y prueba se observa, un porcentaje de estimación correcta del 80.4%, ver Figuras 7 y 8.

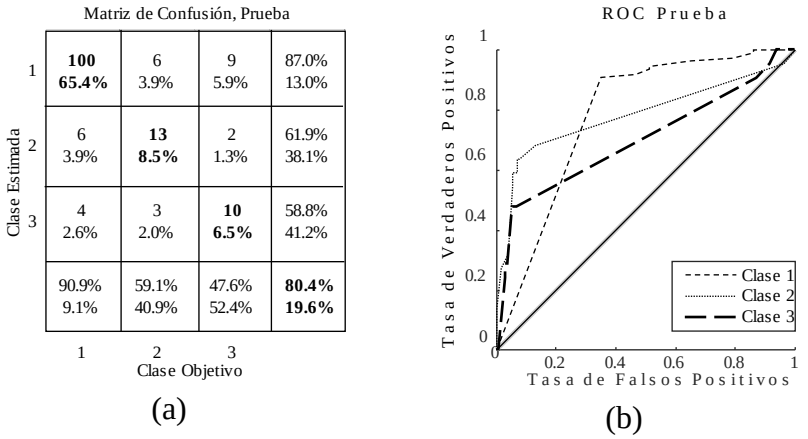

Fig. 8. Matriz de confusión en la fase de prueba. Porcentaje de predicción igual al de validación. (b) Gráfico de "Receiver operating characteristic".

5. Conclusiones y trabajo futuro

El ensamble de clasificadores tuvo un mejor desempeño al evaluarse sobre el conjunto de prueba, 80.4%, que los clasificadores individuales, que oscilan entre 62% y 65.5% o del clasificador que consideraba a todos los factores, 58.62%. Por lo que a través de estos resultados se puede concluir que el ensamble de clasificadores construido con la red neuronal resultó mejor durante la fase de prueba.

Se toman los resultados en la fase de prueba porque es cuando el clasificador se somete a datos considerados durante la fase de entrenamiento, siendo un mejor reflejo de su capacidad para clasificar.

Como trabajo futuro queda evaluar con otros métodos de ensamblaje, como son: el voto mayoritario y la tabla de decisión, entre otros, para evaluar si se mejora el desempeño.

También analizar la ponderación que le otorga la red neuronal del ensamble para identificar al clasificador que más se acercó a la clase correcta, y así examinar la estructura del árbol correspondiente, con el fin de detectar los factores educativos para la clase excelente y generar estrategias para fortalecerlos.

Referencias

1. Zhi-Hua, Z.: Ensemble methods: Foundations and Algorithms. CRC Press, Taylor & Francis Group (2012)

2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), pp. 37–54 (1996)
3. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. Morgan Kaufmann Publishers (2005)
4. WEKA 3: Data Mining Software in Java Homepage. https://www.cs.waikato.ac.nz/ml/weka/ (2016)
5. Singh, Y., Chanuhan, A.: Neural Networks in Data Mining. Journal of Theorical & Applied Information Technology, 5(1), pp.37–42 (2009)
6. Orallo, J., Ramírez, M., Ferri, C.: Introducción a la Minería de Datos. Pearson Education, (2008)
7. Khasawneh, K., Ozsoy, M., Ghazaleh, N., Ponomarev, D.: EnsembleHMD: Accurate Hardware Malware Detectors with Specialized Ensemble Classifiers. IEEE Transactions on Dependable and Secure Computing, pp. 10 (2018)
8. Yan, Y., Yang, H., Wang, H.: Two simple and effective ensemble classifiers for twitter sentiment analysis. Computing Conference 2017, pp. 1386–1393 (2017)
9. Vogado, L., Veras, R., Andrade, A., Araujo, F., Silva, R., Aires, K.: Diagnosing Leukemia in Blood Smear Images Using an Ensemble of Classifiers and Pre-Trained Convolutional Neural Networks. 30th (SIBGRAPI) Conference on Graphics, Patterns and Images, pp. 367–373, Niteroi (2017)
10. Hestenes, M., Stiefel, E.: Methods of Conjugate Gradients for Solving Linear Systems. Journal of Research of the National Bureau of Standards, 49(6), pp 409–436 (1952)