Modelo Rava basado en el análisis comparativo de algoritmos de minería de datos

Huarote Zegarra Raul Eduardo[1], Julca Flores Janett Deisy[1], Castañeda Hilario Aradiel[2], and Jhonatan Isaac Vargas Huaman[3]
[1]Universidad Nacional Tecnológica de Lima Sur, Perú, rhuarote@untels.edu.pe, janettjulca@gmail.com
[2]Universidad Nacional del Callao, Perú, aradiel2006@gmail.com
[3]Universidad Privada del Norte, Perú, jhonatan.vargas@upn.edu.pe

***Abstract*– This research proposes an algorithm called RAVA, named the 4 Data Mining algorithms analyzed qualitatively and quantitatively: Neural Network, Decision Trees, Nearest Neighbor (KNN) and A priori.**

It is worth mentioning that the design of this research is as Sampiero Hernandez not experimental descriptive, based on the use of some evaluation criteria for algorithms such as: precision, clarity, usefulness, adaptability, ease of implementation, speed, sensitivity to noise, and qualitative comparison regarding the calculation complexity is a function of time and space suggested by a sample composed of 5 experts in the analysis of data mining algorithms. This allowed us to identify strengths and weaknesses of the four selected algorithms, resulting in the data mining algorithm RAVA, which has less computational complexity in time and space, observing that the algorithm execution time is lower for RAVA the same amount of data used in the neural network algorithm, the Apriori algorithm, the algorithm of decision tree and nearest neighbor algorithm, on the other hand, the neural network algorithm has a higher time than the other algorithms.

***Keywords*-- Data Mining Algorithms, Neural Network, Decision Trees, Nearest Neighbor (KNN), Apriori.**

I. Introducción

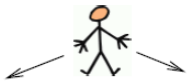
Por ser la Minería de Datos una herramienta que día a día cobra importancia en diversos entornos de nuestra sociedad, desde el académico, pasando por el comercial, hasta el de investigación y desarrollo, este presente trabajo de investigación constituye una oportunidad de uso de esta herramienta que ofrece la tecnología.

El presente trabajo de investigación muestra las principales diferencias entre cuatro algoritmos de Minería de Datos: Redes Neuronales, Árboles De Decisión, Vecino Más Cercano y Algoritmo Apriori y propone el diseño de un algoritmo RAVA, para esto se tienen a disposición los siguientes recursos: Primero, un conjunto de datos con información de ventas de productos del Minimarket Palacios, ubicado en la Urbanización Covicorti, Provincia de Trujillo. Segundo, la implementación de cada uno de los algoritmos seleccionados de Minería de Datos y del algoritmo RAVA, que permitirá la ejecución de los algoritmos seleccionados y la posterior comparación de los resultados obtenidos para cada uno de ellos.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

II. Algoritmos de Minería de datos

Los algoritmos de clasificación de Minería de Datos, tienen como objetivo clasificar algunos objetos en un número finito de clases, en función a sus propiedades u características (Atributos), evidentemente las características elegidas dependen del problema (clasificación) a tratar:



Edad	Ingresos
Astigmatismo	Deudas
Ratio de lagrímeo	Propiedades
Miopía	.....
Tipo de lentillas	Conceder crédito

Fig. 1 Ejemplo a clasificar.

Los algoritmos seleccionados de Minería de Datos, sirvieron como base para el diseño de un nuevo algoritmo RAVA, con el cual la información obtenida permitió resolver los problemas del Minimarket Palacios en el cual se necesitaba determinar la posibilidad del éxito al comercializar un nuevo producto.

En esta sección se presenta los 4 algoritmos de Minería de Datos: Red Neuronal, Árbol decisión, K-NN y Apriori, una descripción de cada uno de ellos, así como sus tipos, sus respectivos algoritmos y sus aplicaciones en la actualidad

A. Redes neuronales

Según Minsky (1951), las redes neuronales artificiales (RNA) han emergido como una potente herramienta para el modelado estadístico orientada principalmente al reconocimiento de patrones –tanto en la vertiente de clasificación como de predicción.

Las RNA poseen una serie de características admirables, tales como la habilidad para procesar datos con ruido o incompletos, la alta tolerancia a fallos que permite a la red operar satisfactoriamente con neuronas o conexiones dañadas y la capacidad de responder en tiempo real debido a su paralelismo inherente.

Las Redes Neuronales Artificiales (RNAs) son la implementación en hardware y/o software de modelos matemáticos idealizados de las neuronas biológicas. Las neuronas artificiales son interconectadas unas a otras y son distribuidas en capas de tal forma que emulan en forma simple la estructura neuronal de un cerebro. Cada modelo de neurona es capaz de realizar algún tipo de procesamiento a partir de estímulos de entrada y ofrecer una respuesta, por lo que las RNA en conjunto funcionan como redes de computación paralelas y distribuidas similares a los sistemas cerebrales biológicos.

B. Árbol de decisión

Según Rossiter (1997), un árbol de decisión es un conjunto de condiciones (reglas) organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas.

C. KNN (K-NEAREST NEIGHBOR)

Según Rossiter (1997), un árbol de decisión es un conjunto de condiciones (reglas) organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas.

Las técnicas de vecinos más cercanos (NN, Nearest Neighbours) basan su criterio de aprendizaje en la hipótesis de que los miembros de una población suelen compartir propiedades y características con los individuos que los rodean, de modo que es posible obtener información descriptiva de un individuo mediante la observación de sus vecinos más cercanos.

Los fundamentos de la clasificación por vecindad fueron establecidos por E. Fix y J. L. Hodges a principio de los años 50. Sin embargo, no fue hasta 1967 cuando T. M. Cover y P. E. Hart enuncian formalmente la regla del vecino más cercano y la desarrollan como herramienta de clasificación de patrones. Desde entonces, este algoritmo se ha convertido en uno de los métodos de clasificación más usados. [Cost & Salzberg, 1993].

La regla de clasificación NN se resume básicamente en el siguiente enunciado: Sea D = $\{e_{1,\ldots,}e_n\}$ un conjunto de datos con N *ejemplos etiquetados*, donde cada ejemplo e_i contiene m atributos ($e_{i1},..,e_{im}$), pertenecientes al espacio métrico E^m, y una clase $C_i \in \{C_1,\ldots,C_d\}$. La clasificación de un nuevo ejemplo e' cumple que

$$\mathbf{e' \rightarrow C_i \Leftrightarrow \forall_{j \neq} \, i.d(e', e_i) < d(e', e_j)}$$

donde $e' \rightarrow C_i$ indica la asignación de la etiqueta de clase Ci al ejemplo e'; y d expresa una distancia definida en el espacio m-dimensional E^m.

Así, un ejemplo es etiquetado con la clase de su vecino más cercano según la métrica definida por la distancia d. La elección de esta métrica es primordial, ya que determina qué significa más cercano. La aplicación de métricas distintas sobre un mismo conjunto de entrenamiento puede producir resultados diferentes. Sin embargo, no existe una definición previa que indique si una métrica es buena o no. Esto implica que es el experto quien debe seleccionar la medida de distancia más adecuada.

La regla NN puede generalizarse calculando los k vecinos más cercanos y asignando la clase mayoritaria entre esos vecinos. Tal generalización se denomina k–NN. Este algoritmo necesita la especificación apriori de k, que determina el número de vecinos que se tendrán en cuenta para la predicción. Al igual que la métrica, la selección de un k adecuado es un aspecto determinante.

El problema de la elección del k ha sido ampliamente estudiado en la bibliografía. Existen diversos métodos para la estimación de k. [Wettschereck & Dietterich, 1993].Otros autores han abordado el problema incorporando pesos a los distintos vecinos para mitigar los efectos de la elección de un k inadecuado. Otras alternativas [Riquelme, Ferrer, & Aguilar, 2001]intentan determinar el comportamiento de k en el espacio de características para obtener un patrón que determine a priori cuál es el número de vecinos más adecuado para clasificar un ejemplo concreto dependiendo de los valores de sus atributos.

En un estudio más recientes, F. J. Ferrer desarrollan un algoritmo de clasificación NN no parametrizado que adapta localmente el valor k. El algoritmo k–NN se engloba dentro de las denominadas técnicas de aprendizaje perezoso (*lazy learning*), ya que no genera una estructura de conocimiento que modele la información inherente del conjunto de entrenamiento, sino que el propio conjunto de datos representa el modelo. Cada vez que se necesita clasificar un nuevo ejemplo, el algoritmo recorre el conjunto de entrenamiento para obtener los k vecinos y predecir su clase. Esto hace que el algoritmo sea computacionalmente costoso tanto en tiempo, ya que necesita recorrer la totalidad de los ejemplos en cada predicción, como en espacio, por la necesidad de mantener almacenado todo el conjunto de entrenamiento. Pese a los numerosos inconvenientes respecto a la eficiencia (coste computacional) y la eficacia (elección de la métrica y el k adecuados), k–NN tiene en general un buen comportamiento.

En la Figura 04 se muestra un ejemplo de aplicación del algoritmo K-NN, Si se toma K=1, el elemento más cercano es circulo y se toma K=7 se clasifica en un cuadrado.

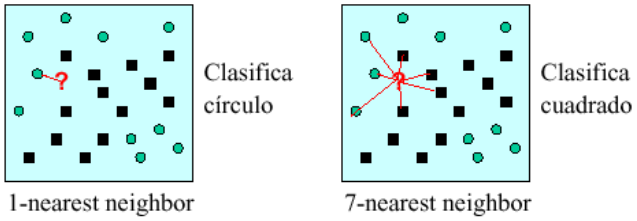


Fig. 2 Ejemplo de método de vecino más cercano.

D. Algoritmo Apriori

Según Rossiter (1997), un árbol de decisión es un conjunto de condiciones (reglas) organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas.

Según Agrawal (1993), éste algoritmo tiene como objetivo obtener itemsets (conjuntos de valores que se repiten) de un determinado tamaño, para combinarlos en reglas de asociación, resultando este algoritmo muy eficiente para grandes volúmenes de datos, sin embargo para ciertos datos de entrada, los resultados intermedios consumen gran cantidad de recursos (memoria).

Las reglas de asociación muestran una correlación estadística entre la ocurrencia de ciertos atributos, en una tabla dentro de una Base de Datos.

Ejemplo:

Defecto_Ocular = Miopía usa_Lentes = SI

La regla general: $X_1, X_2 \ldots X_n \Rightarrow Y$

Donde X_i, Y son atributos de una transacción dada.

III. Resultados

Resultados de la Comparación cualitativa y cuantitativa de los Algoritmos seleccionados de Minería De Datos Vs algoritmo RAVA.

Al realizar las pruebas correspondientes en los Algoritmos seleccionados de Minería De Datos y el algoritmo RAVA, se presenta las siguientes fortalezas/debilidades:

TABLA I

Cuadro comparativo de Fortalezas (F) y Debilidades (D) de los algoritmos seleccionados de Minería de Datos con el algoritmo RAVA

Algoritmos / Fortalezas/Debilidades	**Red neuronal**		**Árbol de decisión**		**Vecino más cercano**		**Apriori**		**RAVA**	
	F	**D**	**F**	**D**	**F**	**D**	**F**	**D**	**F**	**D**
Preciso		X	X			X		X	X	
Claro		X	X		X		X		X	
Útil	X		X		X		X		X	
Adaptable		X		X	X		X		X	
Fácil de implementar	X			X	X			X	X	
Rápido		X	X			X		X	X	
Sensible al Ruido		X		X		X		X		X

El algoritmo RAVA, es preciso pues no presenta errores, además es claro y útil siendo entendible para la solución de problemas y puede trabajar con clases u objetos creados por el usuario considerado una gran estrategia, es adaptable y es fácil de implementar pues puede codificarse en cualquier lenguaje de programación sin embargo para este presente trabajo de investigación se uso el lenguaje de programación de C++. Es rápido y no es sensible al ruido, pues los datos no pueden presentar errores en el valor de un atributo.

Se puede observar que el tiempo de ejecución, del algoritmo Redes neuronales, para la cantidad de datos que emplea es más alto que el de los otros algoritmos de Minería de Datos. Así mismo se aprecia que el tiempo de ejecución promedio del algoritmo RAVA es de 0.0440 segundos, menor que de todos los otros algoritmos, y con una variabilidad de 0.03043 segundos; además los algoritmos de árbol de decisión y vecino más cercano, tienen promedios de ejecución bajos, sin embargo su variabilidad de comportamiento es mayor a al algoritmo RAVA, siendo su variabilidad respectivamente 0.03826 y 0.07538.

Gráfico N° 02: Tiempo medio de ejecución de los algoritmos de Minería de Datos

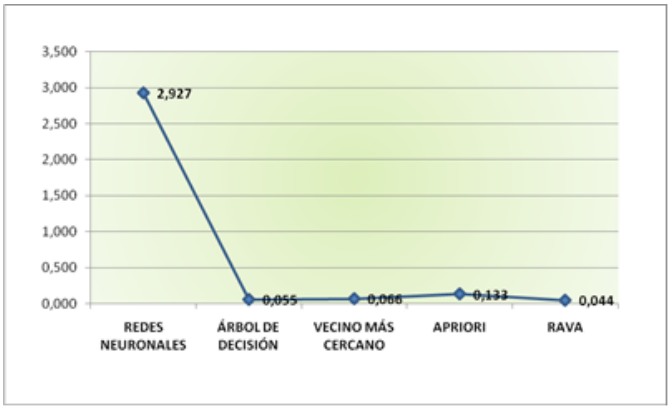


Fig. 3 Lista de Cotejos del análisis de algoritmos de Minería De Datos. 2011.

IV. Conclusiones

El análisis cuantitativo demostró que el tiempo de ejecución promedio del algoritmo RAVA es de 0.0440 segundos, menor que de todos los otros algoritmos, y con una variabilidad de 0.03043 segundos; además los algoritmos de árbol de decisión y vecino más cercano, tienen promedios de ejecución bajos, sin embargo su variabilidad de comportamiento es mayor a al algoritmo RAVA, siendo su variabilidad respectivamente 0.03826 y 0.07538 y respecto al espacio utilizado se concluyó que usa una menor cantidad de memoria.

· El análisis cualitativo de las características de los algoritmos seleccionados de Minería De Datos concluyo que la principal fortaleza es la utilidad y la principal debilidad es la rapidez conforme se aprecia en la Tabla 02.

· El algoritmo de Minería de Datos RAVA, producto del análisis comparativo de algoritmos posee las principales fortalezas: es preciso, claro, útil, adaptable, fácil de implementar, rápido y no es sensible al ruido se aprecia en la Tabla 04.

· El algoritmo de Minería de Datos RAVA fue validado por un grupo de expertos en Análisis de algoritmos de Minería de Datos, obteniendo: 80% afirman que el uso de los recursos tecnológicos para el algoritmo RAVA es muy bueno, el 20% aseguran que es bueno, en tanto que ningún experto afirman que los recursos tecnológicos para el algoritmo RAVA sea regular, malo o muy malo; además el 60% de los expertos afirman que la utilidad del algoritmo RAVA es bueno, el 40% aseguran que es muy bueno, y ningún experto afirman que la utilidad del algoritmo RAVA sea regular, malo o muy malo; además el 60% de los expertos afirman que la adaptabilidad del algoritmo RAVA es bueno, el 20% aseguran que es muy bueno, en tanto que el 20% afirman que la adaptabilidad es regular y ningún experto afirman que la adaptabilidad del algoritmo RAVA sea malo o muy malo; el 60% de los expertos afirman que la implementación usada en el algoritmo RAVA es regular, el 40% aseguran que es malo, en tanto que ningún experto afirman que la implementación sea muy malo, muy bueno y bueno; el 60% de los expertos afirman que la rapidez del algoritmo RAVA es bueno, el 20% aseguran que es muy bueno, en tanto que el 20% afirman que la rapidez es regular y ningún experto afirman que la rapidez del algoritmo RAVA sea malo o muy malo; el 60% de los expertos afirman que la sensibilidad al ruido del algoritmo RAVA es bueno, el 40% aseguran que es muy bueno, en tanto que ningún experto afirman que la sensibilidad al ruido del algoritmo RAVA sea regular, malo o muy malo.

AGRADECIMIENTO

Agradecimiento a todos los participantes de la presente investigación, con el fin de cubrir una necesidad de encontrar un modelo de minería de datos, para poder extraer una información valiosa.

REFERENCES

[1] R. Agrawal, R. Srikant, Q, Vu, Mining association rules with item constraints, in: The Third International Conference on Knowledge Discovery in Databases and Data Mining, 1997, pp. 67–73.

[2] C. Perez, D. Santín. (2007) “Mineria de datos, tecnicas y herramientas” Y. (2017). *Paraninfo Cengage Learning*, Madrid. España. 1-4.

[3] Hormozi, Amir & Giles, Stacy. (2004). Data Mining: A Competitive Weapon for Banking and Retail Industries. IS Management. 21. 62-71. 10.1201/1078/44118.21.2.20040301/80423.9.

[4] Hu, Zhenjiang & Chin, Wei-Ngan & Takeichi, Masato. (2000). Calculating a New Data Mining Algorithm for Market Basket Analysis.. The Journal of Functional and Logic Programming [electronic only]. 2001. 169-184.

[5] Aldehuela M. (2005). *Analisis comparativo entre metodos estadisticos y de mineria de datos*. España.

[6] Wu, X., Kumar, V., Ross Quinlan, J. et al. Top 10 algorithms in data mining. Knowl Inf Syst 14, 1–37 (2008).

[7] J. Gemma, “MINERÍA DE DATOS: ¿QUÉ RELACIÓN TIENE CON EL BIG DATA?” Recuperado el 02 de 08 del 2018 de https://br.escueladenegociosydireccion.com/business/big-data/la-mineria-de-datos-en-el-big-data/

[8] Rivas E., “¿Qué es el Data Mining o minería de datos?” Recuperado el 08 de 01 del 2018 de https://www.iebschool.com/blog/data-mining-mineria-datos-big-data/

[9] Rivas-Asanza, Wilmer & Mazon-Olivo, Bertha & Mejia, Fernando. (2018). Capítulo 1: Generalidades de las redes neuronales artificiales.

[10] Oliveira Colabone, R., Ferrari, A. L., Da Silva Vecchia, F. A., & Bruno Tech, A. R. (2015). Application of Artificial Neural Networks.

[11] Barrientos, Rocío Erandi et al (2009). Árboles de decisión como herramienta en el diagnóstico Médico. México. Revista Médica de la Universidad Veracruzana, No. 2.

[12] J. Pilar. (2012). “Herramientas para la Gestión y la Toma de Decisiones”, Editorial Hanne.

[13] E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951).

[14] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory, 13*, 21-27.

[15] Amat. J. (2017). “Arboles de prediccion: bagging, random forest boosting y c5.0” VECTORITCGROUP.

[16] L. Rouhiainen.,“Inteligencia artificial, 101 cosas que debes saber hoy sobre nuestro futuro”, Alienta Editorial., España, Noviembre 2018.