Recibido: 14/oct/2016
Aprobado: 04/mayo/2017

ARTÍCULO ORIGINAL

Métodos de minería de datos ligados a la inteligencia artificial aplicables a riesgo crediticio

Data mining methods linked to artificial intelligence -applicable to credit risk

Patricia Jimbo Santana

Universidad Central del Ecuador – Facultad de Ciencias Administrativas- Carrera de Contabilidad y Auditoría. Quito, Ecuador
e-mail:prjimbo@uce.edu.ec

Augusto Villa Monte

Instituto de Investigación en Informática LIDI, Universidad Nacional de la Plata, 50 y 120, La Plata, Buenos Aires, Argentina
e-mail: avillamonte@lidi.info.unlp.edu.ar

Laura Lanzarini

Instituto de Investigación en Informática LIDI, Universidad Nacional de la Plata, 50 y 120, La Plata, Buenos Aires, Argentina
e-mail: laural@lidi.info.unlp.edu.ar

Aurelio f. Bariviera

Departament of Business, Universitat Rovira i Virgili, Avenida de la Universitat, 1 Reus, Spain
e-mail: aurelio.fernandez@urv.cat

Resumen

Cuando las instituciones financieras seleccionan, apropiadamente, a sus clientes disminuyen su riesgo de crédito. Los bancos utilizan diversas metodologías con la finalidad de clasificar a sus clientes de acuerdo al riesgo de impago que presentan; para esto se analiza un conjunto de variables personales, así como la situación financiera del cliente que es sujeto de crédito. El análisis y procesamiento exhaustivo de la información del cliente lleva bastante tiempo, una de las causas es que los datos a analizar no son homogéneos. En este trabajo se presenta un método alternativo capaz de construir, a partir de la información disponible, un conjunto de reglas, de clasificación con tres características principales: precisión adecuada, baja cardinalidad y facilidad de interpretación. Esto último está dado por el uso de un número reducido de atributos en la conformación del antecedente. Esta característica, sumada a la baja cardinalidad del conjunto de reglas permite distinguir patrones sumamente útiles a la hora de comprender las relaciones entre los datos y tomar decisiones. Cuando se trata de decidir el otorgamiento de créditos, resulta sumamente útil contar con una herramienta de este tipo. Mientras más simple sea el modelo, menor será la cantidad de características del sujeto de crédito que deben ser analizadas, por lo que las decisiones pueden tomarse con mayor rapidez; esto permite que el método resulte atractivo para los oficiales de crédito en las instituciones financieras, ya que se puede dar una respuesta al solicitante del crédito en menor tiempo logrando una ventaja competitiva. La metodología pro-puesta ha sido aplicada a dos bases de datos conocidas en la literatura y a dos bases de datos reales de entidades financieras del Ecuador, una Cooperativa de Ahorro y Crédito y un Banco que otorgan diferentes tipos de cré-ditos y tienen agencias en las regiones costa, sierra y oriente. Los resultados obtenidos han sido satisfactorios. Finalmente se exponen las conclusiones y se sugieren futuras líneas de investigación.

Palabras clave: riesgo crediticio; reglas de clasificación; redes neuronales competitivas; optimización mediante cúmulos de partículas

Abstract: The Financial institutions, when properly selecting their clients, reduce their credit risk, banks use different methodologies in order to classify their clients according to the default risk they present; For this we analyze a set of personal variables as well as the financial situation of the client that is subject to

credit. The exhaustive analysis and processing of customer information takes a long time, one reason being that the data to be analyzed are not homogeneous. This paper presents an alternative method capable of constructing, from the available information, a set of classification rules with three main characteristics: adequate accuracy, low cardinality and ease of interpretation. The latter is given by the use of a reduced number of attributes in the conformation of the antecedent. This feature, added to the low cardinality of the set of rules allows to distinguish very useful patterns in the understanding of the relations between the data and to make decisions. When it comes to deciding the granting of credits, it is extremely useful to have a tool of this type. The simpler the model, the smaller the number of characteristics of the subject of credit that must be analyzed so that decisions can be taken more quickly. This allows the method to be attractive to credit officers in financial institutions, since It´s possible to give a response to the applicant of the credit in less time obtaining a competitive advantage. The proposed methodology has been applied to two databases known in the literature and to two real databases of Ecuadorian financial institutions, a Savings and Credit Cooperative and a Bank that grant different types of loans and have agencies in the coast, Sierra and oriente. The results obtained have been satisfactory. Finally the conclusions are presented and future lines of research are suggested.

Key words: credit risk; classification rules; competitive neural networks; optimization by particle clusters

Introducción

La economía en el mundo actual conlleva a que las personas soliciten créditos, con diferente finalidad como son los créditos productivo, comercial, de consumo, vivienda, inmobiliario, microcrédito e incluso crédito de inversión pública. Esto lleva a que las instituciones financieras analicen una gran cantidad de variables microeconómicas que le permitan realizar el análisis del sujeto de crédito, y así poder dar una respuesta sobre el crédito solicitado, y de acuerdo a su capacidad financiera establecer su forma de pago.

Por otro lado, en la actualidad, gracias al avance tecnológico, numerosos procesos registran de manera automática sus operaciones dando lugar a grandes repositorios de información histórica. Este registro contiene no sólo información proveniente de distinto tipo de observaciones sino el resultado de decisiones tomadas oportunamente, lo que motiva el interés por aprender a partir de situaciones pasadas buscando identificar los criterios utilizados.

La Minería de Datos ha dado respuesta a este problema a través de distintas técnicas que modelizan la información disponible, sin la necesidad de disponer de una hipótesis previa, por lo que al realizar el análisis de las variables crediticias se puede ofrecer respuestas al análisis financiero.

El objetivo de este trabajo es modelizar la información de riesgo crediticio a través de reglas de clasificación. La identificación adecuada de las características más relevantes será de gran ayuda para la toma de decisiones por parte del analista financiero, debido a que le va a llevar menos tiempo en su análisis, dando una respuesta al sujeto de crédito en el menor tiempo posible.

Para medir el desempeño del método propuesto se analizan distintas soluciones considerando especialmente la simplicidad del modelo en lo que se refiere a:

La cantidad de reglas: cuanto menor sea la cardinalidad del conjunto de reglas mejor será el modelo obtenido.

La longitud promedio del antecedente de las reglas: cuanto menos condiciones se utilicen para formar el antecedente de cada regla, más fácil será la interpretación del modelo.

Una regla de asociación es una expresión de la forma

SI condic1 ENTONCES condic2

donde ambas condiciones son conjunciones de proposiciones de la forma (atributo=valor) y cuya única restricción es que los atributos que intervienen en el antecedente de la regla no formen parte del consecuente. Cuando el conjunto de reglas de asociación presenta en el consecuente el mismo atributo se dice que se trata de un conjunto de reglas de clasificación [2] [18].

Este artículo presenta un método de obtención de reglas de clasificación que combina una red neuronal con una técnica de optimización. El énfasis está puesto en alcanzar una buena cobertura utilizando un número reducido de reglas.

La sección 2 describe brevemente algunos trabajos relacionados, la sección 3 el método propuesto, la sección 4 expone los resultados obtenidos y la sección 5 resume las conclusiones y describe algunas líneas de trabajo futuras.

Trabajos relacionados

En la década de 1960, el desarrollo de los mercados de capitales en Estados Unidos, mostró la necesidad

de utilizar modelos más científicos para evaluar la fuerza corporativa económica. En consecuencia, se desarrolló el primer modelo zscore por Altman [3]. Un trabajo de relevamiento de técnicas aplicadas en el ámbito financiero publicado hacia fines de la década del '90 [4], no da aún testimonio explícito de la aplicación de modelos de hazard rate (tasa de riesgo) [20] o verosimilitud parcial [21]. Sí se deja constancia del uso de las técnicas estadísticas probit y logit conjuntamente con técnicas de transición de estados y otras denominadas "derivación de probabilidades de default de tipo actuarial" asociadas al default pasado de bonos. Iniciado el nuevo milenio se fueron dando a conocer desarrollos específicos para la aplicación del análisis de supervivencia a la medición del riesgo de crédito [5] [14][15]. En las últimas décadas, ha habido un aumento en el crédito de consumo. En nuestro entorno las cooperativas de ahorro y crédito son consideradas como una industria en crecimiento, no sólo ha habido un auge en socios con tarjeta de crédito, especialmente en las economías emergentes, sino también un aumento de pequeños créditos de consumo, por ejemplo, es muy común en las economías que las familias compren electrodomésticos con tarjetas de crédito a través de cuotas; en varios países, es habitual la asociación de una tienda de electrodomésticos con una institución financiera, con el fin de ofrecer a los clientes una línea de crédito rápida. La existencia de tal instrumento financiero ayuda a aumentar las ventas. Esta asociación genera un conflicto de intereses, por un lado, la tienda de electrodomésticos quiere vender productos a todos los clientes; por lo que le interesa promover una política de crédito atractiva. Por otro lado, la entidad financiera quiere maximizar los ingresos procedentes de los créditos, que conducen a una estricta vigilancia de las pérdidas en los préstamos otorgados. El objetivo es que existan políticas transparentes entre las tiendas que ofertan los electrodomésticos y las instituciones financieras También se presenta el caso de las instituciones financieras que otorgan directamente crédito de consumo o microcrédito, cuyo interés es la minimización del riesgo. Una forma de desarrollar dicha política es la construcción de reglas objetivas con el fin de decidir conceder o denegar una solicitud de crédito.

El utilizar técnicas computacionales inteligentes produce mejores resultados, estas técnicas, sin ser exhaustivas, incluyen redes neuronales artificiales, teoría de conjuntos difusos, árboles de decisión, máquinas de vectores soporte, algoritmos genéticos, entre otros. En lo que se refiere a las redes neuronales, existen distintas arquitecturas según el tipo de problema a resolver. Estas arquitecturas incluyen modelos populares, tales como son las redes de propagación hacia atrás, los mapas auto-organizativos o SOM (Self-organizing maps) y el aprendizaje de cuantificación vectorial o LVQ (Learning Vector Quantization). Teoría de conjuntos difusos, desarrollado a partir del trabajo seminal por Zadeh [19] resulta muy útil en casos tales como la clasificación de crédito, donde los límites no se encuentran bien definidos. Los datos pueden también estructurarse en forma de árboles, con sus respectivas ramas, donde el objetivo es poner a prueba los atributos de cada rama del árbol, también se pueden utilizar las máquinas de vectores de soporte las mismas que según el tipo de función discriminante que se utilice permiten construir modelos lineales y no lineales sumamente potentes. Los algoritmos genéticos así como la optimización mediante cúmulos de partículas, son técnicas de optimización poblacionales inspiradas en distintos procesos biológicos.

Si se tiene como objetivo obtener reglas de asociación se puede utilizar el método a priori [1] o alguna de sus variantes. Este método se encarga de identificar los conjuntos de atributos que son más comunes y luego los combina para obtener un conjunto de reglas. Hay variantes del método a priori, que se encargan de reducir el tiempo de cálculo.

Si se trabaja con reglas de clasificación, la literatura contiene distintos métodos de construcción basados en árboles como el C4.5 [13] o en árboles recortados como el método PART [6], en cualquiera de los casos, lo fundamental es obtener un conjunto de reglas que cubra los ejemplos cumpliendo con una cota de error preestablecida. Los métodos de construcción de reglas a partir de árboles son partitivos y se basan en distintas métricas de los atributos a fin de estimar su capacidad de cobertura.

Metodología

Este artículo presenta una metodología que puede ser considerada híbrida basada en la combinación de cúmulos de partículas (PSO - Particle Swarm Optimization) con redes neuronales competitivas. Estas últimas son utilizadas para comenzar la búsqueda en posiciones prometedoras. Si bien existen métodos de obtención de reglas utilizando PSO [17], cuando se opera sobre atributos nominales es preciso contar con suficientes ejemplos como para cubrir todas las zonas del espacio de búsqueda y esto no siempre es factible. El resultado es una pobre inicialización de la población lo que da lugar a la convergencia prematura. Para resolver este problema y a la vez reducir el tiempo de obtención, se comparó el rendimiento de varios métodos que combinan población fija y variable, PSO inicia con dos redes neuronales competitivas LVQ (Learning Vector Quantization) y SOM (Self Organizing Maps). Existen en la literatura métodos que utilizan PSO como forma de determinar la cantidad óptima de neuronas competitivas a utilizar en la red, como por ejemplo [7]. Esta no es la propuesta de este trabajo ya que la técnica de optimización se utiliza aquí para identificar las características más representativas que formarán los antecedentes de las reglas.

Learning Vector Quantization

Learning Vector Quantization (LVQ) es un algoritmo de clasificación supervisado basado en centroides o prototipos [Kohonen 1990], este algoritmo puede ser interpretado como una red neuronal competitiva formada por tres capas; la primera capa es sólo de entrada, la segunda es donde se realiza la competencia, la tercera capa que es la de salida es la encargada de realizar la clasificación. Cada neurona de la capa competitiva lleva asociado un vector numérico de igual dimensión que los ejemplos de entrada y una etiqueta que indica la clase a la cual va a representar, estos vectores son los que al finalizar el proceso adaptativo contienen la información de los centroides o prototipos de la clasificación. Existen distintas versiones del algoritmo de entrenamiento, a continuación se describe la que utilizamos en este artículo.

Al iniciar el algoritmo, debe indicarse la cantidad K de centroides que van a ser utilizados, esto permite definir la arquitectura de la red donde la cantidad de entradas y salidas están definidas por el problema a resolver.

Los centroides se inicializan tomando K ejemplos aleatorios, luego se ingresa cada uno de los ejemplos y se procede a adaptar la posición de los centroides, luego de lo cual se determina el centroide más cercano al ejemplo de turno utilizando una medida de distancia preestablecida. Como se trata de un proceso supervisado es posible determinar si el ejemplo y el centroide correspoden o no a la misma clase, si el centroide y el ejemplo pertenecen a la misma clase, se "acerca" el centroide al ejemplo con el objetivo de fortalecer la representación, si, por el contrario, las clases son distintas, se "aleja" el centroide. Estos movimientos se realizan utilizando un factor o velocidad de adaptación, el mismo que permite ponderar el paso que se va a realizar.

Este proceso se repite hasta que las modificaciones que se vayan a realizar sean menores, a un umbral establecido anteriormente o hasta que los ejemplos se identifiquen con los mismos centroides en dos iteraciones consecutivas, lo que primero que ocurra.

Como una variante en la implementación realizada en este artículo, también se considera al segundo centroide más cercano, siempre que la clase a la que pertenezca sea distinta a la del ejemplo analizado, y se encuentre a una distancia inferior a 1.2 veces la distancia del primero, debido al factor de inercia que fue establecido anteriormente y al "alejamiento" aplicado.

Pueden consultarse distintas variantes de LVQ en [10]

Self-Organization Maps

La red neuronal SOM (Self-Organizing Maps) fue definida por Kohonen en 1982 [10], su función principal es la de agrupar toda la información disponible, se caracteriza por su capacidad de preservar la topología de los datos de entrada. Al igual que LVQ, se trata de una técnica de agrupamiento partitiva, ya que asocia cada ejemplo a un vector promedio o un centroide. Sin embargo incorpora la noción de vecindad entre los centroides permitiendo que agrupamientos similares se encuentren más próximos dentro de la arquitectura. Esta característica no existe en LVQ. Por esta razón, es utilizada como herramienta de visualización así como para reducir el número de dimensiones del espacio de entrada. Esta se puede representar como una estructura de dos capas: la capa de entrada, cuya función es sólo para permitir ingresar información a la red, y la capa competitiva, la misma que es responsable de la tarea de agrupamiento. Las neuronas que forman esta segunda capa están conectadas y tienen la capacidad de identificar el número de "saltos" o conexiones que los separan de cada una de las otras neuronas en este nivel, cada neurona competitiva está asociada a un vector de peso o centroides representada por los valores de los arcos que llegan a esta neurona de la capa de entrada. Por lo tanto, la red SOM interactúa con dos estructuras de información: una en relación con los centroides vinculados a las neuronas competitivas, y la otra es responsable de determinar la proximidad alrededor de las neuronas. Este estilo, a diferencia de otros métodos tal como el método K-means [8], ofrece información adicional sobre los clusters, ya que las neuronas que están muy cerca dentro de la arquitectura pueden representar grupos similares en el espacio de datos de entrada.

Obteniendo reglas de clasificación con Particle Swarm Optimization (PSO)

La optimización mediante cúmulo de partículas o PSO (Particle Swarm Optimization) es una metaheurística poblacional propuesta por Kennedy y Eberhart [9] donde cada individuo de la población, denominado partícula, representa una posible solución del problema y se adapta siguiendo tres factores: su conocimiento sobre el entorno (su valor de aptitud), su conocimiento histórico o experiencias anteriores (su memoria) y el conocimiento histórico o experiencias anteriores de los individuos situados en su vecindario (su conocimiento social).

La obtención de reglas de clasificación utilizando PSO, capaces de operar sobre atributos nominales y numéricos, requiere de una combinación de los métodos citados anteriormente ya que es preciso decir cuá-

les serán los atributos que formarán parte del antecedente y cuál es el valor o rango de valores que podrán tomar (combinación de discreto y continuo).

Por tratarse de una técnica poblacional, debe analizarse la información requerida en cada individuo de la población, hay que decidir entre representar una única regla o el conjunto completo por individuo y elegir el esquema de representación de cada regla. De acuerdo a los objetivos planteados en este trabajo se siguió el enfoque Iterative Rule Learning (IRL) [16] en el que cada individuo representa una única regla y la solución del problema se construye a partir de los mejores individuos obtenidos en una secuencia de ejecuciones. Por ello, la utilización de este enfoque implica que la técnica poblacional se aplique de manera iterativa hasta lograr la cobertura deseada obteniendo una única regla en cada iteración: el mejor individuo de la población. Además se ha decidido utilizar una representación de longitud fija donde sólo se codificará el antecedente de la regla y dado el enfoque adoptado, se efectuará un proceso iterativo asociando todos los individuos de la población con una clase predeterminada lo cual no requiere de la codificación del consecuente.

Método propuesto para obtención de reglas

Las reglas se obtienen a través de un proceso iterativo que analiza los ejemplos no cubiertos de cada clase comenzando por las más numerosas. Cada vez que se obtiene una regla, los ejemplos cubiertos correctamente por dicha regla son retirados del conjunto de datos de entrada. El proceso continúa hasta lograr cubrir todos los ejemplos o hasta que la cantidad de ejemplos no cubiertos de cada clase se encuentre por debajo del soporte mínimo establecido o hasta que se hayan realizado la máxima cantidad de intentos por obtener una regla, lo que ocurra primero. Es importante tener en cuenta que, dado que los ejemplos son retirados del conjunto de datos de entrada a medida que son cubiertos por las reglas, las mismas constituyen una lista de clasificación. Es decir que, para clasificar un ejemplo nuevo, las reglas deben ser aplicadas en el orden en que fueron obtenidas y el ejemplo será clasificado con la clase correspondiente al consecuente de la primera regla cuyo antecedente se verifique para el ejemplo en cuestión.

Debido a que las redes neuronales sólo operan con datos numéricos, los atributos nominales son repre-sentados mediante una codificación ficticia o "dummy" que utiliza tantos dígitos binarios como opcio-nes distintas posea dicho atributo nominal, además, antes de iniciar el entrenamiento, cada dimensión correspondiente a un atributo numérico es escalada linealmente en [0,1]; la medida de similitud utilizada es distancia euclídea, una vez finalizado el entrenamiento, cada centroide contendrá aproximadamente el promedio de los ejemplos que representa.

Para obtener cada una de las reglas se determina, en primer lugar, cual es la clase correspondiente al consecuente, buscando obtener reglas con soporte alto, el método propuesto comenzará a analizar primero las clases que posean un mayor número de ejemplos no cubiertos. El soporte mínimo que debe cumplir una regla es proporcional a la cantidad de ejemplos no cubiertos de la clase al momento en que fue obtenida, es decir, que el soporte mínimo requerido para cada clase disminuye a lo largo de las iteraciones, a medida que los ejemplos de la clase correspondiente se van cubriendo. De esta forma, es de esperar que las primeras reglas posean mayor soporte que las últimas.

En la Figura 1 se muestra el pseudocódigo del método propuesto.

Para más detalles consultar [11] y [12].

Entrenar la red neuronal competitiva utilizando todos los ejemplos de entrenamiento.

Calcular el soporte mínimo para cada clase.

Mientras (no se alcance el criterio de terminación)

Elegir la clase con mayor nro.de ejemplos no cubiertos

Construir una población reducida de individuos a partir de los centroides

Evolucionar la población utilizando PSO según lo visto en la sección 4

Obtener la mejor regla de la población Si(la regla cumple con el soporte y la confianza pedidos) entonces

Agregar la regla al conjunto de reglas

Considerar como cubiertos los ejemplos correctamente clasificados por la regla anterior.

Recalcular el soporte mínimo para esta clase.

Fin Si

Fin mientras

Figura 1. Pseudocódigo del método propuesto

Resultados

Se realizaron pruebas con dos bases de datos del repositorio UCI y dos bases de datos reales de instituciones ecuatorianas, para esta dos últimas se analizaron solicitudes de crédito así como operaciones de crédito concedidas, con los siguientes atributos: el estado; fecha de la solicitud; destino del crédito; provincia; monto requerido; monto autorizado; propósito del

crédito; efectivo con el que cuenta el cliente, cuentas bancarias, inversiones, otros activos, pasivos y sueldo del solicitante; fecha de la verificación de la información; fecha de autorización; fecha de la aprobación/ negación; cuentas bancarias, inversiones, otros activos, pasivos y sueldo del cónyuge del solicitante. En el caso de que el solicitante sea un pequeño negocio los datos solicitados son los ingresos y los gastos del negocio. Las solicitudes de crédito pueden ser negadas o aceptadas, en caso de ser aceptadas se realiza una nueva clasificación entre los créditos que fueron cancelados sin ninguna novedad y los que tienen algún retraso en la recuperación de la inversión. A su vez, los créditos vencidos se clasifican, de acuerdo con las políticas crediticias entre los que tienen menos de 90 días de retraso, y los que tienen más de 90 días de retraso (inicio de acciones legales), que son aquellos se consideran que pasan a cartera vencida.

El uso de los datos descritos anteriormente, compara el rendimiento de varios métodos que combinan dos tipos de PSO, uno de población fija y otro de población variable, inicializadas con dos redes neuronales competitivas diferentes: LVQ y SOM. Se comparan estas soluciones con los métodos C4.5 y PART. La manera de encontrar reglas de clasificación en los métodos propuestos y de control son diferentes, C4.5 es un árbol podado cuyas ramas son excluyentes entre sí y permiten clasificar los ejemplos. PART da como resultado una lista de reglas equivalentes a las generadas por el método de clasificación propuesto, pero en una forma determinista. El funcionamiento de PART se basa en la construcción de árboles parciales. Cada árbol se crea de una manera similar a la propuesta para C4.5 pero durante el proceso se calculan errores de construcción de cada rama, estos errores determinan la poda del árbol.

El método propuesto utiliza en su algoritmo valores aleatorios que hacen que el movimiento de la partícula no sea excesivamente determinista, como en el caso de PART, la característica más importante de los resultados obtenidos, es la combinación de un algoritmo de búsqueda de atributos con una red neuronal competitiva, lo que da un conjunto de reglas con una cardinalidad significativamente baja (menor cantidad de reglas). Aunque los algoritmos de partición proporcionan una mayor precisión, esto se consigue a través de un mayor número de reglas, lo que hace más difícil la comprensión. De hecho, la diferencia en la precisión entre ambos tipos de métodos está dentro del intervalo de 1 a 3 puntos porcentuales. La precisión de la clasificación basada en PSO es muy buena y es comparable a los otros métodos, sin embargo, en relación el número de reglas es entre10 y 20 veces mayor en los métodos de partición.

Por lo que se puede considerar un tipo de compensación entre la exactitud y la sencillez de las reglas, el objetivo es que el oficial de crédito pueda responder rápido con la mayor exactitud verificando el menor número de reglas, por lo que consideramos que este método es una buena alternativa.

Tabla 1. Resultados con base de datos Australiana – repositorio UCI

https://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval)

Método		Verdadero +	Verdadero -	Falso +	Falso -	Precision	#reglas	Long.del antecedente
SOM + PSO	**Media**	**0.4257**	**0.4333**	**0.1097**	**0.0309**	**0.8590**	**3.0167**	**1.3711**
	desv	0.0154	0.0103	0.0069	0.0066	0.0099	0.0461	0.1922
SOM + varPSO	**Media**	**0.4183**	**0.4391**	**0.1071**	**0.0351**	**0.8574**	**3.0000**	**1.5178**
	desv	0.0132	0.0158	0.0130	0.0077	0.0104	0.0000	0.1085
LVQ + PSO	**Media**	**0.4201**	**0.4414**	**0.1079**	**0.0306**	**0.8614**	**3.0000**	**1.2667**
	desv	0.0179	0.0172	0.0093	0.0065	0.0105	0.0000	0.1207
LVQ + varPSO	**Media**	**0.4199**	**0.4382**	**0.1054**	**0.0363**	**0.8582**	**3.0000**	**1.5578**
	desv	0.0179	0.0172	0.0075	0.0073	0.0092	0.0000	0.1336
C4.5	**Media**	**0.3910**	**0.4618**	**0.0847**	**0.0625**	**0.8528**	**18.2200**	**4.8394**
	desv	0.0121	0.0063	0.0066	0.0120	0.0124	2.0825	0.2810
PART	**Media**	**0.3564**	**0.3906**	**0.1562**	**0.0969**	**0.7469**	**33.3433**	**2.4926**
	desv	0.0136	0.0288	0.0289	0.0134	0.0292	1.5793	0.0934

Tabla 2. Resultados con base de datos Alemana – repositorio UCI

https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

Método		Verdadero +	Verdadero -	Falso +	Falso -	Precision	#reglas	Long.del antecedente
SOM + PSO	**Media**	**0.1026**	**0.5993**	**0.0984**	**0.1994**	**0.7019**	**8.4400**	**2.1619**
	desv	0.0123	0.0183	0.0149	0.0183	0.0153	0.6009	0.1415
SOM + varPSO	**Media**	**0.1046**	**0.5954**	**0.1034**	**0.1965**	**0.7000**	**8.0233**	**2.0464**
	desv	0.0115	0.0135	0.0110	0.0134	0.0162	0.6745	0.1030
LVQ + PSO	**Media**	**0.0999**	**0.5997**	**0.1017**	**0.1986**	**0.6996**	**8.7767**	**2.1802**
	desv	0.0151	0.0171	0.0136	0.0153	0.0133	0.7224	0.1075
LVQ + varPSO	**Media**	**0.1089**	**0.5973**	**0.1057**	**0.1880**	**0.7063**	**8.8867**	**2.0884**
	desv	0.0100	0.0129	0.0103	0.0116	0.0109	0.4918	0.0960
C4.5	**Media**	**0.1219**	**0.5894**	**0.1106**	**0.1781**	**0.7113**	**86.4600**	**5.6267**
	desv	0.0069	0.0070	0.0070	0.0069	0.0079	4.0788	0.1382
PART	**Media**	0.1404	0.4385	0.1687	0.2258	0.6967	70.9133	3.0138
	desv	0.0120	0.0091	0.0135	0.0170	0.0139	2.1575	0.0561

Tabla 3. Resultados con base de datos de Cooperativa de Ahorro y Crédito del Ecuador que se encuentra en el segmento 2 dentro de la SuperIntendencia de Economía Popular y Solidaria, activos mayor a 20'000.000,00 hasta 80'000.000,00

Método		Verdadero +	Verdadero -	Falso +	Falso -	Precision	#reglas	Long.del antecedente
SOM + PSO	**Media**	**0.6242**	**0.1601**	**0.1253**	**0.0898**	**0.7844**	**3.7867**	**1.6375**
	desv	0.0069	0.0062	0.0057	0.0059	0.0059	0.2980	0.2151
SOM + varPSO	**Media**	**0.6014**	**0.1914**	**0.0947**	**0.1125**	**0.7928**	**4.1533**	**1.6953**
	desv	0.0052	0.0059	0.0057	0.0047	0.0030	0.2801	0.0867
LVQ + PSO	**Media**	**0.6227**	**0.1671**	**0.1191**	**0.0910**	**0.7899**	**3.2933**	**1.4021**
	desv	0.0048	0.0055	0.0051	0.0039	0.0031	0.1837	0.1066
LVQ + varPSO	**Media**	**0.6029**	**0.1902**	**0.0956**	**0.1114**	**0.7930**	**4.3733**	**1.6553**
	desv	0.0056	0.0055	0.0054	0.0053	0.0025	0.2625	0.0567
C4.5	**Media**	**0.6320**	**0.1786**	**0.1075**	**0.0819**	**0.8106**	**114.2600**	**9.6762**
	desv	0.0014	0.0013	0.0013	0.0013	0.0011	6.0543	0.1144
PART	**Media**	**0.6229**	**0.1825**	**0.1036**	**0.0910**	**0.8054**	**42.3567**	**4.6956**
	desv	0.0065	0.0064	0.0064	0.0065	0.0023	2.1661	0.0880

Tabla 4. Resultados con base de datos de Institución Financiera

Banco del Ecuador que se encarga de dar crédito de consumo productivo y no productivo

Método		Verdadero +	Verdadero -	Falso +	Falso -	Precision	#reglas	Long.del antecedente
SOM + PSO	**Media**	**0.0457**	**0.8863**	**0.0228**	**0.0451**	**0.9320**	**4.3967**	**5.4962**
	desv	0.0058	0.0047	0.0045	0.0056	0.0050	0.4895	0.4221
SOM + varPSO	**Media**	**0.0609**	**0.8870**	**0.0210**	**0.0300**	**0.9480**	**3.8600**	**2.8940**
	desv	0.0029	0.0059	0.0051	0.0030	0.0063	0.2415	0.3532
LVQ + PSO	**Media**	**0.0482**	**0.8870**	**0.0219**	**0.0428**	**0.9352**	**4.6067**	**5.9166**
	desv	0.0052	0.0056	0.0055	0.0049	0.0054	0.4193	0.2771
LVQ + varPSO	**Media**	**0.0565**	**0.8882**	**0.0198**	**0.0346**	**0.9447**	**3.8533**	**3.1013**
	desv	0.0043	0.0062	0.0056	0.0043	0.0056	0.2921	0.3395
C4.5	**Media**	**0.0762**	**0.9017**	**0.0073**	**0.0148**	**0.9779**	**153.5733**	**11.2349**
	desv	0.0003	0.0003	0.0003	0.0003	0.0003	5.1687	0.1565
PART	**Media**	**0.0457**	**0.8863**	**0.0228**	**0.0451**	**0.9320**	**4.3967**	**5.4962**
	desv	0.0058	0.0047	0.0045	0.0056	0.0050	0.4895	0.4221

Conclusiones

Se ha presentado un nuevo método de obtención de reglas de clasificación aplicadas al análisis de riesgo crediticio basado en la combinación de PSO (optimización por cumulo de partículas) y redes neuronales competitivas. Su aplicación se ha realizado con dos bases de datos de créditos reales de una Cooperativa de Ahorro y Crédito así como de un Banco del mercado Ecuatoriano, y de dos bases de datos públicas presentes en el repositorio UCI (UC Irvine Machine Learning Repository) ha sido satisfactoria. Las mediciones realizadas permiten afirmar que el método propuesto permite reducir significativamente el número de reglas que se requiere y alcanzar un nivel aceptable de precisión.

Es importante destacar que el objetivo de este trabajo de investigación es lograr un modelo intuitivo para el scoring de crédito que permita una precisión comparable a los modelos de referencia populares. Los resultados obtenidos sugieren que la simplificación de las reglas de decisión genera transparencia en la puntuación de crédito, lo que podría mejorar la reputación de las instituciones financieras. El modelo obtenido en este trabajo de investigación ha logrado alcanzar una buena precisión utilizando un número reducido de reglas, las mismas que son simples en su interpretación y representan baja cardinalidad (menor número de atributos analizados del sujeto de crédito) disminuyendo el tiempo en el análisis de la solicitud de crédito, ayudando a la toma de decisiones de los oficiales de crédito, lo que presenta una ventaja competitiva frente a las instituciones financieras ya que el tiempo en dar la respuesta a una solicitud de crédito se reduce considerablemente.

En futuras líneas de investigación se debe considerar el incorporar el análisis conjunto de variables microeconómicas con variables macroeconómicas, que permitan un modelo más simple manteniendo una precisión adecuada.

Referencias bibliográficas

Aggarwal C. (2015). Data Mining: The Textbook. Springer Publishing Company, Incorporated.

Agrawal, R., Srikant, R. (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, pp. 487–499.

Altman, E.I., (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), pp.589–609.

Altman, E.I., and Sounders, A. (1998). credit risk measurement: developments over the last 20 years, Journal of Banking and Finance 21, 1721-1742.

Duffie, D., Singleton K. J. (2003). credit risk: pricing, measurement, and management. princeton university press. ISBN 0-691-09046-7.

Frank, E., Witten, I. H., (1998). Generating accurate rule sets without global optimization. In: proceedings of the fifteenth international conference on machine learning, ICML '98., pp. 144–151.

Hung, C. & Huang, L. (2010). Extracting rules from optimal clusters of self-organizing maps. In second international conference on Computer Modeling and Simulation. ICCMS '10. pp. 382–386

Kennedy, J. & Eberhart, R. (1995). Particle swarm optimization. In proceedings of IEEE International Conference on Neural Networks. pp. 1942–1948 vol.4.

Kohonen, T. (2012). Self-Organizing Maps. Volume 30, springer series in information sciences. Springer, Heidelberg.

Lanzarini, L., Villa Monte, A., Aquino, G., De Giusti, A. (2015). Obtaining classification rules using lvqPSO Advances in swarm and computational intelligence. Lecture notes in computer science. Vol 6433, 183-193. Heidelberg: Springer-Verlag Berlin.

Lanzarini, L., Villa-Monte, A., Ronchetti, F. (2015). SOM+PSO. A novel method to obtain classification rules. Journal of computer science & technology (JCS&T), 15(1), 15-22.

MacQueen, J. B. (1967), Some methods for classification and analysis of multivariate observations," in Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, pp. 281,297.

Nanda,A. & Shaked, M. (2001). The hazard rate and the reversed hazard rate orders, with applications to order statistics, pp.853–864.

Quinlan, J.R. (1993). C4.5: programs for machine learning, Morgan Kaufmann Publishers.

Ríos Insua, M. (1984). On the hierarchical models and their relationship with the decision problem with partial information a priori.

Roszbach, K. (2003): Bank lending policy, credit scoring and the survival of loans. sveriges risksbank working paper series 154.

Saunders, A., Allen L. (2002). Credit risk measurement: new approaches to value at risk and other paradigms, 2nd Edition. John Wiley & Sons, Inc. ISBN: 978-0-471-27476-6.

Venturini G., (1993). Sia: A supervised inductive algorithm with genetic search for learning attributes based concepts," in Machine Learning: ECML-93, ser. lecture notes in computer science, P. Brazdil, Ed. Springer Berlin Heidelberg, vol. 667, pp. 280–296.

Wang, Z., Sun, X. & Zhang, D. (2007). A PSO-Based classification rule mining algorithm. In D.-S. Huang, L. Heutte, & M. Loog, eds. advanced intelligent computing theories and applications. with aspects of artificial intelligence: third international conference on intelligent computing, ICIC 2007, Qingdao, China, August 21-24, 2007. Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 377–384.

Witten, I.H., Eibe, F. & Hall, M.A. (2011). Data mining practical machine learning tools and techniques 3rd. ed., San Francisco, CA: Morgan Kaufmann Publishers Inc.

Zadeh, L.A. (1965). Fuzzy sets. Information and Control, 8(3), pp.338–353.