Software para clasificación/predicción de datos

Software for Data Classification/Prediction

Jorge Enrique Rodríguez Rodríguez
Magíster en Ingeniería de Sistemas. Especialista en Ingeniería de Software. Especialista en Diseño y Construcción de Soluciones Telemáticas. Ingeniero de Sistemas. Docente investigador de la Universidad Distrital Francisco José de Caldas. Director del Grupo de Investigación en Inteligencia Artificial de la misma universidad.
Correo electrónico: jrodri@udistrital.edu.co

Clasificación del artículo: investigación

Fecha de recepción: 8 de febrero de 2007

Fecha de aceptación: 19 de julio de 2007

Palabras clave: minería de datos, clasificación, predicción, red neuronal, algoritmo de retropropagación.

Key words: data mining, data classification, prediction, neural networks, backpropagation algorithm.

Resumen

En este artículo se muestra el proceso de desarrollo y aplicación de un software para la clasificación de datos (llamado DClass), que hace parte del proyecto de investigación: Desarrollo de Herramientas para Minería de Datos (UDMiner). Se inicia con una introducción. Luego, se especifica sobre la fase diseño de la herramienta; ésta incluye: selección de la técnica, diseño del modelo de clasificación y construcción del software. Posteriormente, se abordan diferentes problemas, a fin de hacer las pruebas correspondientes. Por último, se plantean las conclusiones obtenidas de este avance parcial de investigación.

Abstract

This paper reflects the development and application of software for the Classification of Data (called DClass); it makes part of the research project "Development of Tools for Data Mining - UDMiner". It starts with an introduction, then comes the design phase which includes: selection of the technique, design of the model of classification, and construction of the software. Later on different problems are undertaken in order to make the corresponding tests. Finally the conclusions obtained from this partial advancement of the research are shown.

* * *

1. Introducción

La minería de datos es la exploración y análisis, de forma automática o semiautomática, de grandes volúmenes de datos para el descubrimiento de patrones y reglas significativas [1]. En nuestro medio, específicamente, en Colombia no se ha explorado o aplicado en profundidad. Esto trae como consecuencia la realización de una amplia documentación (consulta de textos, "papers", aplicaciones, etc.) sobre la apropiación de tales elementos en las diferentes organizaciones en las que se emplea la minería de datos. Por lo tanto, este avance parcial de investigación requirió de revisión bibliográfica para especificar un estado del arte, puesto que no se puede negar que exista mucha bibliografía –tanto primaria como secundaria y terciaria– para poder determinar cuáles son las aplicaciones potenciales en nuestro medio.

Ahora bien, es un hecho que diversas organizaciones colombianas no han adoptado tecnologías de la información y las comunicaciones que, por ejemplo, les permita determinar los productos alimenticios que un cliente con cierto perfil social compraría, así como la predicción automática de tendencias y comportamientos, análisis de las ventas de una compañía farmacéutica para reforzar las acciones de marketing en los hospitales y médicos de mayor impacto, prescripción de dietas alimenticias, identificación de mejores clientes para el lanzamiento de una nueva tarjeta de crédito; entre otras aplicaciones en la que la minería de datos juega un papel fundamental.

Por consiguiente, con el desarrollo de este software se está dando un aporte significativo a la empresa nacional, para así solucionar diferentes problemas de clasificación/predicción que éstas puedan presentar. De igual forma, se impulsó la investigación en esta naciente área del conocimiento y se está dando un aporte al diseño y al desarrollo de este tipo de herramientas [2].

El presente artículo se encuentra estructurado en referentes teóricos; diseño del software, iniciando con la justificación de la selección de la técnica empleada en la clasificación/predicción y el modelo de minería de datos planteado, y pruebas y análisis de resultados. Con el desarrollo del software no se pretende competir con herramientas de propósito comercial (por ejemplo, DBMiner, o Microsoft OLE DB para Minería de Datos), sino ofrecer un prototipo de software para ser utilizado en la academia y un conocimiento apropiado. Sin embargo, se puede tomar como base para que determinadas empresas continúen desarrollando esta herramienta y finalmente se pueda llevar a cabo un producto de software final que usen en actividades que crean conveniente para mejorar su competitividad.

2. Referentes teóricos

2.1. Clasificación/predicción

Las bases de datos contienen mucha información oculta que puede ser usada para tomar decisiones inteligentes. La clasificación/predicción es una forma de análisis de datos que puede ser usada para extraer modelos que describan las más importantes clases de datos o para predecir el comportamiento futuro de los mismos. Diversos métodos de clasificación/predicción han sido propuestos por los investigadores en aprendizaje computacional, sistemas expertos, estadística y neurobiología. La mayoría de los algoritmos están residentes en memoria y asumen una pequeña cantidad de datos. Recientemente las investigaciones en minería de datos han aportado nuevos conceptos, a fin de lograr escalabilidad en los métodos de clasificación/predicción; de este modo se ha logrado manejar gran cantidad de datos residentes en disco. Estas técnicas consideran el procesamiento distribuido y paralelo.

2.1.1. ¿Qué es clasificación/predicción?

La clasificación de datos es un proceso de dos pasos: en el primer paso se construye un modelo que describa el conjunto preliminar de clases. El modelo es construido mediante el análisis de los registros ejemplos. Cada registro pertenece a una clase específica conocida, debido a esto, la clasificación se enmarca dentro de lo que se conoce

como aprendizaje supervisado. En contraste, en el aprendizaje no supervisado (también relacionado con clustering), la clase a la que pertenece cada registro es desconocida, y el número de clases por aprender tampoco puede ser conocido. Generalmente, el modelo aprendido es representado en la forma de reglas de clasificación, árboles de decisión o fórmulas matemáticas [3].

En el segundo paso, el modelo es usado para generar la clasificación de datos desconocidos. Luego, se estima la precisión del modelo o clasificador. La precisión de un modelo en un conjunto dado de datos es el porcentaje de ejemplos del conjunto de entrenamiento que fueron correctamente clasificados. Si la precisión del modelo es considerada aceptable, el modelo puede ser usado para clasificar futuros conjuntos de datos para los cuales la etiqueta de la clase es desconocida.

2.1.2. ¿Qué diferencia existe entre clasificación y predicción?

La minería de datos puede ser vista desde dos puntos de vista: predictivo y descriptivo. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o independientes, usando otras variables o campos de las bases de datos, llamadas variables independientes o predictivas [4]. Algunas de las tareas que producen modelos predictivos son la clasificación y la regresión; por consiguiente, la clasificación es una forma de predecir. Los modelos predicitvos tienen el objetivo específico de permitir predecir valores desconocidos de variables de interés, a partir de valores de otras variables [5].

2.2. La Minería de Datos en Colombia

Según la Dirección de Metodología y Producción Estadística en un informe de Medición de las Tecnologías de la Información y las Comunicaciones (Departamento Administrativo Nacional de Estadística) realizada en junio de 2001 en las instituciones del Estado y publicada en julio de 2002, el tipo de software existente en las instituciones es: 30,9%, en programas de usuario final; 27,9%, en programas de uso administrativo; el 15,5%, en manejadores de bases de datos; 10,9%, en herramientas de programación; el faltante 14,8% se refiere a programas de uso administrativo, de producción, de procesos, de minería de datos, bodega de datos y otros (tabla 1).

El 7 de enero de 2003, la BBRITE –una división de Infórmese Ltda.– publicó un artículo sobre la implementación de la segunda fase del proyecto SARC de la Superintendencia Bancaria de Colombia. Ese proyecto fue implementado por LLoyds TSB Bank, que desarrolló los modelos de minería de datos requeridos para atender los requerimientos, tanto para la banca personal como para la banca corporativa.

La empresa UNI/ BASE Ltda. es una empresa profesional de alto nivel en el área de la informática; fue escogida como representante exclusivo para Colombia por la firma Business Objects, que, a su vez, es líder mundial en herramientas integradas para el soporte en la toma de decisiones para el mercado corporativo, con su herramienta Business Objects. Ésta "Provee un conjunto completo de herramientas DSS incluyendo consultas, reportes, procesamiento analítico de información, minería de datos y módulos de administración para los ambientes cliente/ servidor e Internet". En Colombia se implanta en empresas como Suratep, Conavi, y Éxito.

La minería de datos se utiliza en diversas industrias y en diferentes funciones de negocios:

- En empresas de telecomunicaciones, tarjetas de crédito y compañías de seguros utilizan minería de datos para la detección de fraudes, optimización de campañas de marketing, descripción y segmentación de clientes, y predicción de fidelidad de clientes.
- En la industria del comercio se utiliza la minería de datos para diseñar y evaluar campañas de marketing; definir ofertas más apropiadas o recomendaciones de productos a clientes, y predecir riesgo en asignación de créditos a clientes.

Tabla 1. Tecnologías de la información y las comunicaciones 2001

Cuadro 1.4.1.10 2001									
	Tipo de software que existía en la instución								
	Usuario final	Uso administrativo	Producción	Procesos	Herramientas de programación	Manejador de bases de datos	Mineria de datos	Bodega de datos	Otros
Nacional	2,335	2,038	626	328	821	1,173	42	71	122
CVE	1,36	1,95	6,08	8,89	4,96	3,86	23,34	19,39	15,48

Fuente: Encuesta de Tecnologías de la Información y las Comunicaciones, 2001.
Nota 1: Las estimaciones con CVE entre 15.0 y 25.0 se deben manejar con precaución.
Nota 2: Las estimaciones con cve mayores a 25.0 no se presentan.
CVE: Coeficiente de variación estimada (%)

- En la medicina se utiliza la minería de datos para predecir la efectividad de procedimientos quirúrgicos, exámenes médicos y medicamentos.
- En bancos e instituciones financieras, se utiliza para mejorar el rendimiento económico.

3. Selección de la técnica para la clasificación/predicción de datos (redes neuronales artificiales)

Con el avance en las tecnologías de bases de datos, las compañías iniciaron un gran almacenamiento de sus datos históricos. En la década pasada se incrementó el uso de redes neuronales para minar datos a partir de grandes volúmenes de información. La minería de datos basada en redes neuronales artificiales tiene múltiples dominios de aplicación, incluyen: finanzas, comercio electrónico, medicina, análisis de mercadeo, manejo de inventarios, etc. [6].

El campo de las redes neuronales fue originalmente manejado por sicólogos y neurobiólogos que se dedicaron a desarrollar y a evaluar el comportamiento de las neuronas. Una red neuronal es un conjunto de neuronas de entrada/salida conectadas entre sí, en la cual cada conexión tiene un peso asociado. Durante la fase de aprendizaje la red aprende ajustando estos pesos, de tal manera que está en capacidad de predecir la clase a la que pertenecen los ejemplos. El aprendizaje por redes neuronales también es referido como aprendizaje conexionista debido a las conexiones entre las unidades. Las redes neuronales artificiales son una abstracción computacional del modelo neuronal humano [7].

Las ventajas de las redes neuronales incluyen su alta tolerancia al ruido, como también su habilidad para clasificar patrones sobre los cuales ellas no han sido entrenadas. Adicionalmente, diferentes algoritmos se han desarrollado recientemente para la extracción de reglas provenientes de redes neuronales entrenadas. En gran medida, estos factores contribuyen a la utilización de redes neuronales para clasificación en minería de datos.

A continuación se presenta una justificación de por qué se seleccionó esta técnica: por una parte, las grandes cantidades de datos con que cuentan las empresas obligan a buscar métodos para analizar la información que tienen almacenada; de este modo surge la Minería de Datos como un medio para estudiar de una forma más profunda y eficiente los datos recolectados. Por otra parte, para la minería de datos se tienen técnicas como las redes neuronales que son eficientes para los objetivos buscados, pero de compleja comprensión en su funcionamiento.

La claridad de este modelo y su poder de predicción tienen una relación estrecha, que se comporta de manera inversa, como se muestra en la gráfica 1. Se observa que entre más sencilla sea la forma del modelo, más fácil será su comprensión, pero tendrá menor capacidad para encontrar diferencias sutiles o demasiado variadas. Las que mejor rendimiento

tienen, en cuanto exactitud de resultados, son las redes neuronales que tienen la posibilidad de adaptarse a valores indefinidos o incluso ausentes, pero es casi imposible inspeccionarlas, puesto que es como si se tratará de inspeccionar el cerebro de una persona para saber que está pensando. Solamente las predicciones realizadas pueden ser inspeccionadas y visualizadas.

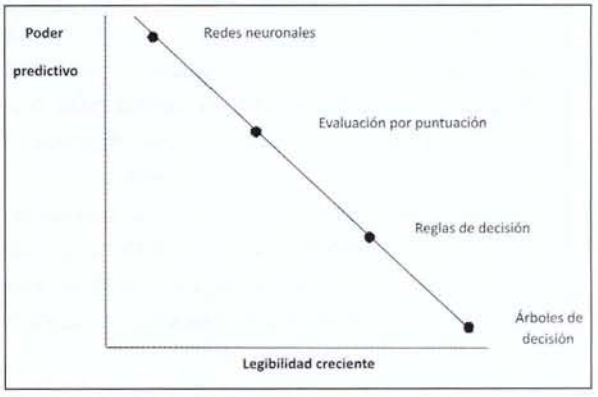
Las redes neuronales son útiles en la minería de datos, debido a su capacidad para modelar datos complejos y multidimensionales. La disponibilidad de datos ha crecido tanto, como la dimensión de los problemas por solucionar, lo que ha limitado muchas técnicas tradicionales, tales como, el análisis manual de los datos y algunos métodos estadísticos; mientras que las redes neuronales ofrecen cualidades como: búsqueda automática de todas las correlaciones posibles entre los hechos clave, un modelado automático de problemas complejos sin el conocimiento "a priori" del nivel de la complejidad, capacidad de extraer los resultados clave más rápidamente que otras técnicas. Las redes neuronales son altamente tolerantes al ruido, al igual que se destaca su habilidad a la hora de clasificar patrones sobre los cuales ellas no han sido entrenadas. Es necesario destacar que recientemente algunos algoritmos han sido diseñados para extraer reglas a partir de una red neuronal entrenada. Aspectos como éstos contribuyen de manera significativa a la utilización de redes neuronales para clasificación de datos.

Otra ventaja del uso de redes neuronales es la parsimonia de la técnica, ya que mediante una red neuronal se puede abordar tanto un problema de clasificación como un problema de regresión, mientras que desde la perspectiva estadística clásica se han necesitando dos modelos tan diferentes como el análisis discriminante y las series temporales. Aunque las redes neuronales son hábiles para encontrar los modelos ocultos en los datos, no revelan directamente sus resultados. El principal problema de las redes neuronales es que su funcionamiento es difícil de entender, esto se debe a factores como:

- Las redes neuronales son un modelo que dificulta la comprensión.
- La relación entre pesos y variables es complicada.
- No permite una comprensión intuitiva de los resultados.

Además, presenta otra desventaja importante que es el tiempo de entrenamiento, y esto se debe a que el error decrece como una potencia del tamaño del entrenamiento. Por último, cabe mencionar que, a menudo, se requiere un significativo procesamiento previo de los datos para adaptarlos a las entradas de la red neuronal.

Es importante acotar que la minería de datos es un proceso que, gracias al apoyo recibido de diferentes tecnologías que han evolucionado actualmente, funciona de manera eficiente y confiable. Esta eficiencia, característica de la minería de datos, es variable dependiendo de la técnica que se utilice en el modelamiento del sistema, así se establece que las redes neuronales son las que obtienen mejores resultados; sin embargo, su dificultad de implementación es mayor. A continuación, se selecciona una topología de red y su tipo de aprendizaje junto con la regla de aprendizaje:



Gráfica 1. Comparación de diferentes algoritmos de minería de datos [8]

- Topología (feedforward): la topología de la red neuronal seleccionada para clasificar y para predecir datos, es una Feedforward. En este tipo de redes, todas las señales neuronales se propagan hacia adelante a través de las capas de la red. No existen conexiones hacia atrás y normalmente tampoco autorrecurrentes, ni laterales, excepto en algunos modelos (las redes feedforward no utilizan conexiones laterales, excepto en los modelos de red propuestos por Kohonen denominados Learning Vector Quantizer y Topology Preserving Map), en las que existen unas conexiones implícitas muy particulares entre las neuronas de la capa de salida.
- Tipo de aprendizaje (supervisado): las actividades de clasificación/predicción son ejemplo de minería de datos dirigida: un objetivo particular es utilizar los datos disponibles para construir un modelo que escriba una determinada variable de interés en función del resto de la información disponible. Los algoritmos supervisados (el aprendizaje supervisado se caracteriza, porque el proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo que determina la respuesta que debería generar la red a partir de una entrada determinada) predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).
- Regla de aprendizaje (backpropagation, también conocidas como retropropagación): las redes con aprendizaje supervisado con una arquitectura de varios niveles utilizan algoritmos de aprendizaje tales como el de backpropagation. El algoritmo backpropagation encuentra un valor mínimo de error (local o global) mediante la aplicación de pasos descendentes y es uno de los más populares; se realiza en dos etapas: al inicio se le da a la red un conjunto de entradas y salidas; como ejemplo, una trama de entrada estimula la primera capa de unidades de la red y se va propagando hacia las unidades de las capas superiores, hasta generar una salida. Esta salida se compara con la salida deseada y se calcula una unidad de error para cada salida. La señal de error se propaga hacia las unidades en las capas inferiores que contribuyeron directamente con la salida, las cuales reciben un fragmento de la señal, proporcional a la contribución de cada unidad a la salida. Con esta nueva información se actualizarán todos los valores de las conexiones en cada nodo, hasta que la red pueda analizar correctamente todas las tramas de entrenamiento, así cada uno de los nodos se especializa en reconocer una característica especifica de la entrada y emite una salida activa sólo si ésta se encuentra presente en la trama de entrada. Gracias a esto la red puede reconocer diferencias sutiles en tramas incompletas o con ruido y ofrecer como salida la mejor aproximación a la salida correcta. Una de las principales razones que conllevaron a la selección de backpropagation es su gran capacidad de autoadaptar los pesos de las neuronas de las capas ocultas para aprender la relación que existe entre un conjunto de patrones dados como ejemplo y sus salidas correspondientes. En el diagrama 1 se resume el algoritmo backpropagation; para mayor detalle de éste, se recomienda ver [9]. En cuanto a la función de activación, el modelo backpropagation requiere el uso de neuronas cuya función de activación sea continua y, por lo tanto, diferenciable. Por lo anterior, la función de activación seleccionada es la sigmoidal.

Aunque la regla backpropagation es la más común en los algoritmos de aprendizaje dentro de las redes neuronales, se han propuesto otros, incluyendo algoritmos más especializados para algunas tareas. Por ejemplo, métodos de redes

neuronales recurrentes, y algoritmos, como una correlación en cascada que altera la estructura y los pesos de la red [10].

4. Especificaciones del software

El modelo utilizado para la clasificación/predicción está basado en redes neuronales artificiales, en el cual se tiene en cuenta: tipo de aprendizaje (supervisado), topología de la red (feedforward), algoritmo de entrenamiento (backpropagation) y función de activación (sigmoide). En la capa de entrada se tiene *n* neuronas, tanto el número de neuronas de entrada como el número de neuronas de salida serán determinados por el usuario del software que desee aplicar minería; este número depende del problema en cuestión; *n* indica el número de entradas. Estas neuronas indican el número de atributos o variables predictoras, en la capa oculta (pueden ser más de una capa oculta) se tienen por defecto *(n + m)/2* neuronas: el número de neuronas de la capa oculta se determina a partir de la experimentación que varía dependiendo del problema en cuestión. Por defecto, este número está dado por el número de neuronas de entrada más el número de neuronas de salida, dividido entre 2. En la capa de salida *m* neuronas cada una de las neuronas de la capa de salida tomará el valor de 0 ó 1, cuando se trate de clasificación, indicando la pertenencia o no pertenencia a determinada clase. Por consiguiente, sólo una neurona de la capa de salida asumirá el valor de 1. También en esta última capa de salida, cada neurona representa una clase. Estos valores son tomados por defecto del archivo preprocesado, los cuales pueden ser modificados por el minero, según lo considere. Por otro lado, se suministra una heurística para determinar el número de neuronas en las capas ocultas, a partir de $d = {}^{mo}/_{w}$, en la cual *m* es el número de patrones de entrenamiento, *o* es el número de salidas de la red, y *w* es número total de pesos de la red [11].

- En cuanto a los parámetros del modelo neuronal, se tiene la tasa de aprendizaje (está dada en el intervalo: $0.0 < \alpha < 1.0$) que permite variar la velocidad de aprendizaje; para tal fin se ofrecen cuatro heurísticas: heurística 1, disminuir gradualmente la tasa luego de que se haya sobre pasado el 50% de las épocas; heurística 2, si luego de un 10% de épocas (en cualquier instante del entrenamiento) el error se incrementa, entonces la tasa de aprendizaje se varia, a fin de bajar el error; heurística 3, cada conexión tiene su propia tasa, definida aleatoriamente [12], y heurística 4, se decrementa la tasa durante todo el proceso de entrenamiento en porciones dadas por la tasa inicial dividida entre el número de épocas. El minero tiene la responsabilidad de seleccionar ninguna o una de estas heurísticas, de acuerdo con el problema abordado. Otro parámetro que agiliza el proceso de aprendizaje, es el momentum (preferiblemente $0.1 \leq momentum \leq 0.25$), este se asume como constante durante todo el proceso de aprendizaje.
- Del número total de patrones se puede seleccionar aleatoriamente un porcentaje para la prueba, a través de la validación cruzada. El proceso de entrenamiento culmina una vez se haya alcanzado el error establecido o el número de épocas. El resultado que se genera, producto del entrenamiento, es una matriz de confusión (para el caso de salidas nominales), en la que se plasma la calidad del aprendizaje, dado por el éxito y el error de la clasificación. Cuando las salidas son continuas, se obtiene una diferencia entre las salidas deseadas y las obtenidas por el modelo para cada patrón. Del mismo modo, para la prueba se genera la matriz de confusión o la diferencia entre la salida deseada y la obtenida. Además, se muestra la gráfica del error de cada época durante la fase de entrenamiento; esto a fin de que el minero pueda determinar puntos críticos durante las fases de entrenamiento.

5. Análisis de pruebas y resultados

Para validar los resultados obtenidos por el software se presentaron diferentes conjuntos de datos (fuente de datos: http://www.ics.uci.edu/~mlearn/MLRepository.html), para cada uno de éstos se realizaron diez corridas, a fin de construir una matriz de

confusión promedio para medir la efectividad del entrenamiento y de la clasificación como tal. En la tabla 2, se muestra una síntesis de los resultados obtenidos por DClass (la efectividad de clasificación mostrada por DClass, corresponde al promedio de diez corridas) frente a los obtenidos por WEKA. En ésta se observa que la efectividad de clasificación de DClass, en general es buena; en los ejemplos mostrados se obtuvo una clasificación correcta por encima del 70% (ocho de los diez ejemplos), sólo en uno (chess) la clasificación es definitivamente deficiente, esto se debe a la incompletitud de los datos. Con WEKA se obtuvo una clasificación del más del 70%, en cuatro de los diez ejemplos.

Con lo anterior, se corrobora el cumplimiento del objetivo de este prototipo de software, el cual consiste en desarrollar una herramienta para la clasificación/predicción de datos.

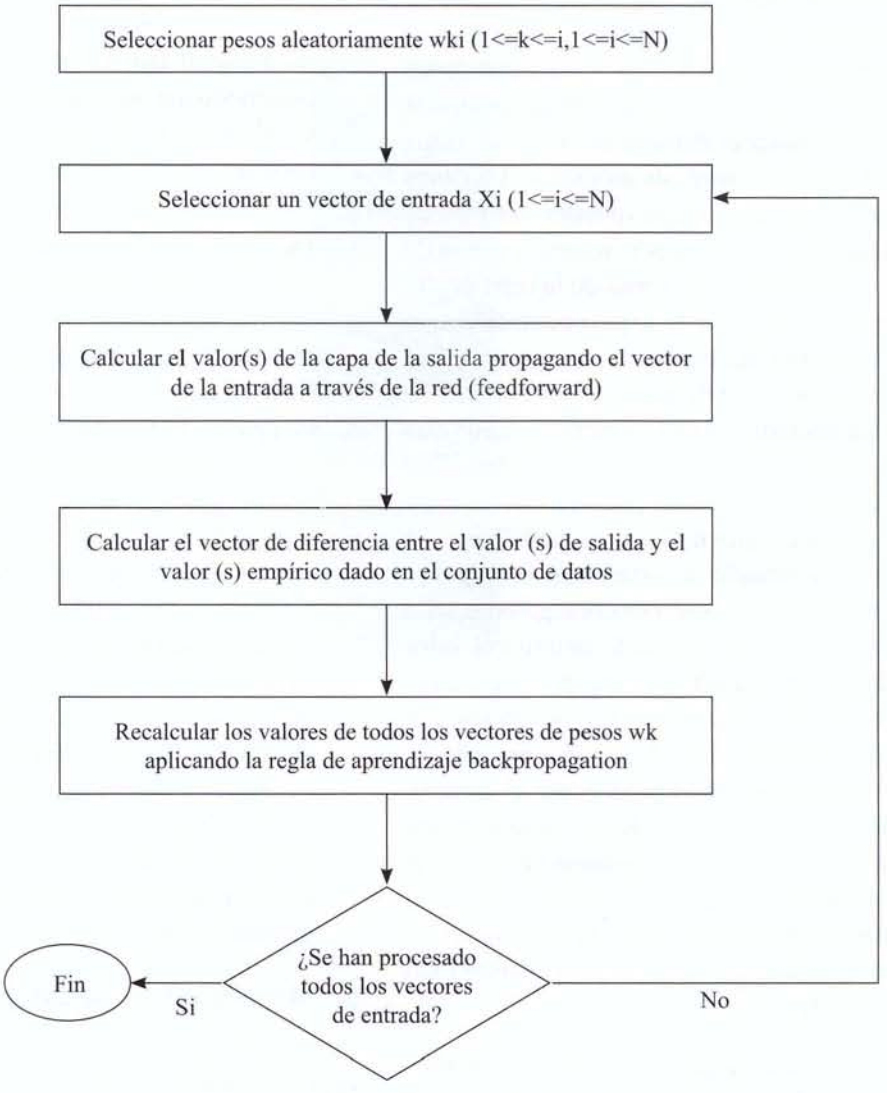


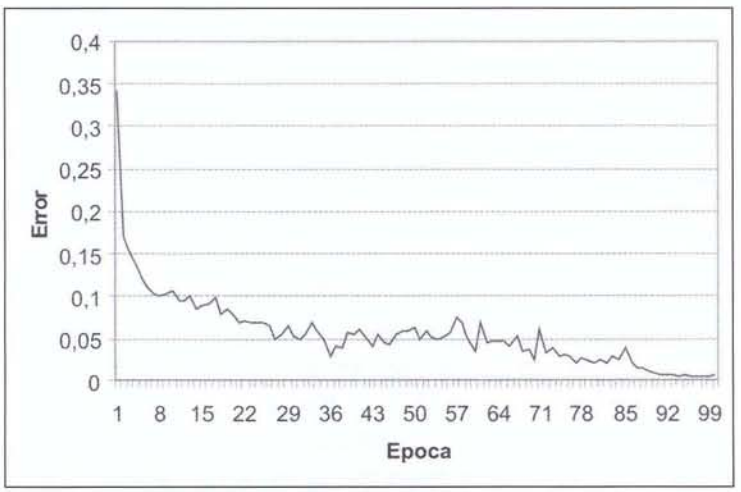
Diagrama 1. Algoritmo de retropropagacion

Por otro lado, para cada ejemplo se probó con diferentes estructuras de redes neuronales, variando el número de neuronas en la capa oculta, el número de capas ocultas, la tasa de aprendizaje, el momentum, el valor inicial de los pesos (aleatoriamente), hasta identificar la estructura de red que generaba menor error. Agregar más de una capa oculta puede llevar a una mejora del desempeño de la red, aunque no es significativo frente al incremento de la complejidad y al tiempo empleado en la clasificación. Incrementar el número de neuronas en la capa oculta mejora el desempeño de la red hasta cierto grado, luego disminuye su desempeño.

Otro aspecto por analizar es la inclusión de diferentes heurísticas para determinar la tasa de aprendizaje (α) durante el entrenamiento de la red; para tal fin se utilizó el conjunto de datos SOYBEAN. En la gráfica 2, se muestra el error obtenido en cada una de las épocas, utilizando una tasa de aprendizaje constante de 0,3. También se observa que se logra una estabilidad luego de aproximadamente el 85% de las épocas; con esta configuración en las épocas iniciales el error decrece relativamente más rápido, que con las demás configuraciones de la tasa.

Tabla 2. Efectividad de entrenamiento y prueba (clasificación) de DClass y WEKA

Ejemplo	DClass		WEKA	
	Entrenamiento	Test	Entrenamiento	Test
Soybean	98,22%	82,15%	97,39%	79,38%
Agaricus	99,68%	99,69%	98,98%	*
Chess	98,42%	3,98%	*	*
House	95,68%	84,11%	?	?
Nourse	100%	80,62%	85%	85,16%
ADN	99,96%	75,82%	73,35%	69,84%
Shuttle	99,76%	80,41%	?	?
Tic-Tac	100%	78,53%	100%	92,91%
Bolsa	97,42%	83,73%	86%	78%
Connect_4	79,64%	57,74%	?	?

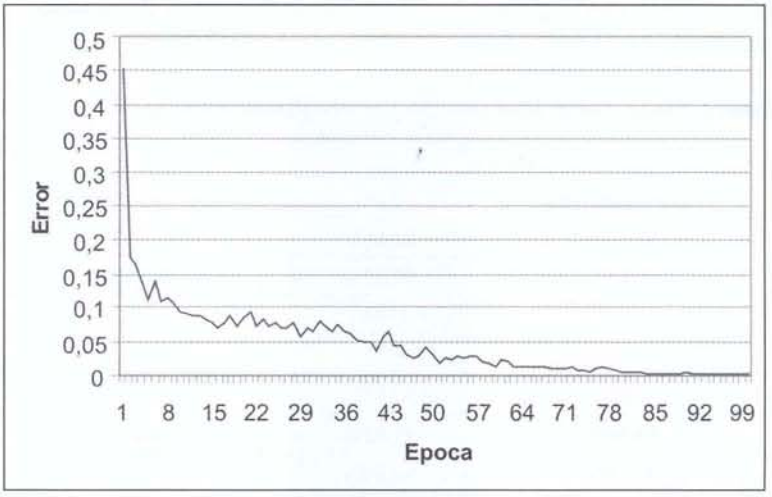


Gráfica 2. Error de entrenamiento con una tasa de aprendizaje constante

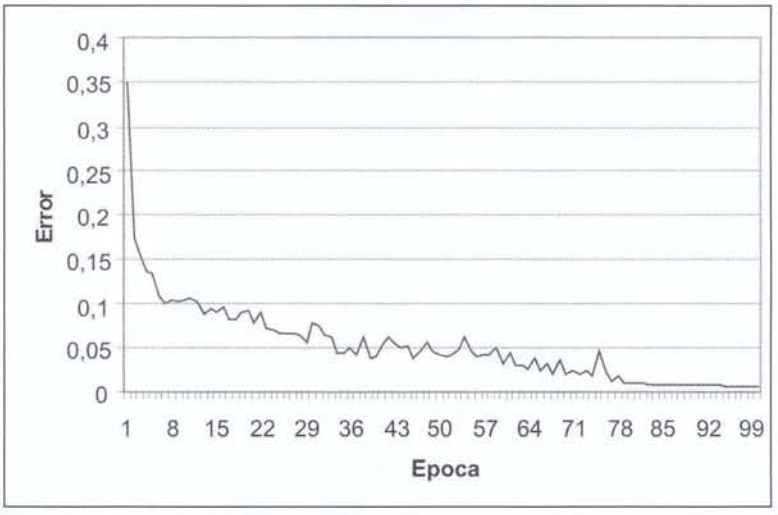
En la gráfica 3 se plasma el comportamiento del error, empleando una heurística que disminuye la tasa de aprendizaje luego de superar el 50% de las épocas. Con esta heurística se obtiene una notable disminución del error durante las primeras épocas, una estabilidad en el entrenamiento más rápido que con las otras heurísticas y un menor error, lo que permite obtener una mejor clasificación; siendo esta heurística la más adecuada –esta heurística también se probó con otros ejemplos, obteniendo un comportamiento del error similar al aquí mostrado.

El del error mostrado en las gráficas 4 y 5 es similar al que se muestra en la gráfica 4, en cuanto al decrecimiento y la estabilidad del error.

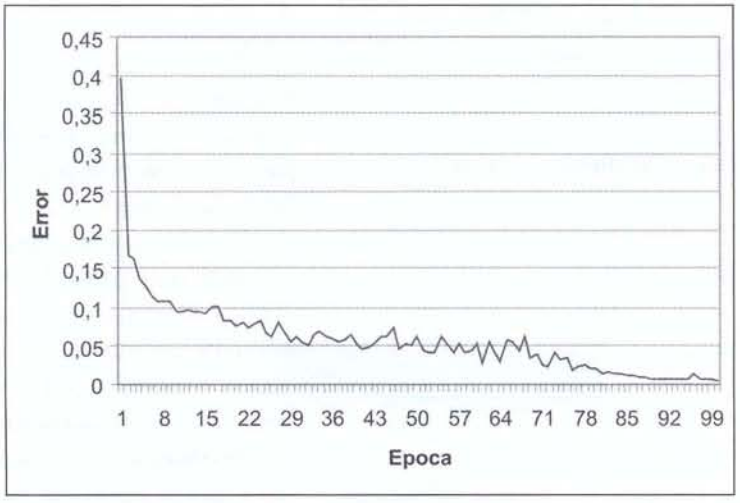
El comportamiento del error durante el entrenamiento con las heurísticas 1 y 4 (gráfica 6) es en términos generales igual; sin embargo se obtiene una mejor clasificación con la heurística 1 (ver tabla 3).



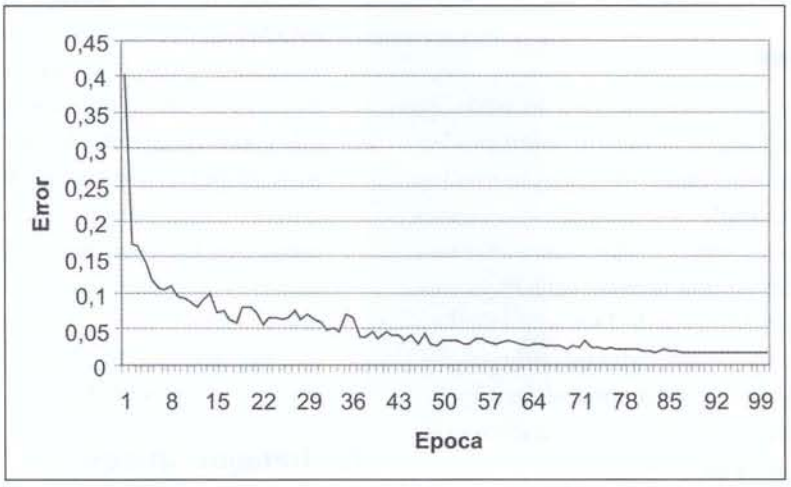
Gráfica 3. Error de entrenamiento con una tasa de aprendizaje disminuida luego del 50% de las épocas del entrenamiento (heurística 1)



Gráfica 4. Error de entrenamiento con una tasa de aprendizaje disminuida, cuando el error sufre un incremento luego de decrecer durante un periodo de épocas (heurística 2)



Gráfica 5. Error de entrenamiento con una tasa de aprendizaje diferente para cada conexión (heurística 3)



Gráfica 6. Error de entrenamiento con una tasa de aprendizaje disminuida a partir de la primera época (heurística 4)

Tabla 3. Efectividad del entrenamiento con las diferentes tasas de aprendizaje

Tasa de aprendizaje (Heurística)	Efectividad del entrenamiento
Tasa constante	77,74%
Heurística 1	82,15%
Heurística 2	80,08%
Heurística 3	79,96%
Heurística 4	80,14%

En la tabla 4 se muestran algunas especificaciones de los conjuntos de datos relacionados.

Tabla 4. Información de los conjuntos de datos

Ejemplo	Instancias	Atributos	Clases
Soybean	683	35	19
Agaricus	8.124	22	2
Chess	28.056	7	18
House	22.784	17	Numérica
Nourse	12.960	9	5
ADN	3.186	181	3
Shuttle	58.000	16	7
Tic-Tac	958	10	2
Bolsa	1.262	12	2
Connect_4	67.557	43	3

6. Conclusiones

- La realización de software para minería de datos, orientado hacia la clasificación/predicción, se terminó con éxito, obteniendo una herramienta que cumple con las especificaciones de desarrollo inicialmente planteadas: diseño e implementación de una técnica (clasificación/predicción) para Minería de Datos, y la selección de un método de aprendizaje computacional, visto como un método aplicado a minería de datos, el cual se puede enmarcar dentro de aprendizaje computacional).
- Toda herramienta de minería de datos sirve como apoyo para la toma de decisiones de las organizaciones, tal es caso de la aplicación en el programa integral de formación laboral, para el que se aplicó; en éste el software se utilizó para la determinación del riesgo de exclusión social. De esta forma, se apoyaron a las directivas del programa integral, para tomar decisiones acerca de los factores que determinan la exclusión social.
- En cuanto a la técnica seleccionada (redes neuronales artificiales), se destaca, entre otras, la parsimonia de ésta, puesto que mediante una red neuronal se puede abordar tanto un problema de clasificación como un problema de predicción, mientras que desde la perspectiva estadística clásica se han necesitado dos modelos tan diferentes como el análisis discriminante (en cuanto a clasificación se refiere) y las series de tiempo (para el caso de predicción). Lo anterior se corrobora en las pruebas, en las que se observa que el algoritmo backpropagation es una buena elección, cuando de clasificar/predecir datos se trata.
- Como aporte, se destaca el desarrollo de una herramienta para minería de datos, pues, según estadísticas del DANE, sólo el 1% del software utilizado o desarrollado en Colombia es de minería de datos; por consiguiente, se concluye que con el desarrollo de esta herramienta, se está aumentando el porcentaje en mención, diferentes universidades y empresas colombianas en los últimos tres años, han hecho desarrollos en el área de minería de datos; las estadísticas presentadas por el DANE no incluyen estos desarrollos, lo que implica, que el 1% planteado por el DANE se haya incrementado.

7. Trabajos futuros

Entrenar el modelo neuronal utilizando Simulated Annealing, a fin de establecer criterios de rendimiento y efectividad con backpropagation. Para aplicar Simulated Annealing, el valor de la tasa de aprendizaje se va disminuyendo gradualmente con el tiempo. Si se cae en un error mínimo local, rápidamente, una corrección considerable, podrá sacarnos de éste [13].

Referencias bibliográficas

[1] Berry, M. (1997) *Data Mining Techniques*. USA: John Wiley & Sons, pp. 5-16.

[2] Rodríguez, J. (2005) *Herramienta software para minería de datos*. Universidad Nacional de Colombia, p. 124.

[3] Han, J. and Kamber, M. (2001) *Data Mining. Concepts and Techniques*. USA: Morgan Kaufmann Publishers, pp. 279-330.

[4] Hernández, J., Ramírez, J., Y Ferri, C. (2004) *Introducción a la minería de datos*. España: Pearson, pp. 3-18.

[5] Hand, D., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining*. USA: The MIT Press, pp. 327-363.

[6] Bozdogan, H. (2004) *Statistical Data Mining and Knowledge Discovery*. USA: Chapman & Hall/CRC, pp. 401-411.

[7] Kantardzic, M. (2003) *Data Mining "Concepts, Models, Methods, and Algorithms"*. USA: IEEE Press Editorial Board, pp. 195-220.

[8] Pmsi. (2002) Data Mining. Disponible en: http://www.pmsi.fr/home-sp.htm

[9] Hilera, J. y Martínez, V. (1995) *Redes Neuronales Artificiales "Fundamentos, Modelos y Aplicaciones"*. España: Addison-Wesley Iberoamericana, pp. 101-154.

[10] Mitchell, T. (1997) *Machine Learning*. USA: McGraw Hill, pp. 81-126.

[11] Peña, L., y Gutiérrez, J. (2005) Redes neuronales artificiales para diagnóstico de neuropatías periféricas focales. Memorias del Congreso Internacional de Inteligencia Computacional, Colombia, pp. 16–24.

[12] Haykin, S. (1999) *Neural Networks "A comprehensive foundation"*. USA: Prentice Hall, pp. 156-252.

[13] Nilsson, N. (1996) Introduction to Machine Learning. USA, pp. 39-68. Disponible en: http://www.cs.stanford.edu