UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Revista Vínculos

https://revistas.udistrital.edu.co/index.php/vinculos/

I+D INVESTIGACIÓN Y DESARROLLO

Algoritmo hibrido de redes neuronales artificiales con recocido simulado para predicción en minería de datos

Hybrid algorithm of artificial neural networks with simulated annealing for prediction in data mining

Roberto Emilio Salas Ruiz[1], Jorge Enrique Rodríguez Rodríguez[2], Claudia Liliana Hernández García[3]

Para citar este articulo: R. E. Salas Ruiz, J. E. Rodríguez Rodríguez, C. L. Hernández García, "Algoritmo hibrido de redes neuronales artificiales con recocido simulado para predicción en minería de datos". *Revista Vínculos*, vol. 17, no. 2, pp. 97-103, julio-diciembre, 2020. https://doi.org/10.14483/2322939X.17232

Recibido: 08-10-2020 / Aprobado: 26-11-2020

Resumen

El presente artículo es un avance del proyecto de investigación titulado "Desarrollo de algoritmos híbridos para minería de datos" y presenta el uso de una red neuronal con el algoritmo del recocido simulado para realizar la predicción de un conjunto de datos de entrenamiento. En primer lugar, se aborda el problema a resolver, el cual está orientado al análisis de las técnicas definidas para el algoritmo hibrido. Luego, se justifica la metodología de investigación (científica descriptiva-exploratoria con enfoque experimental) aplicada. Se realizó la revisión de las técnicas seleccionadas para la técnica híbrida redes neuronales y recocido simulado la cual se aplica a un conjunto de datos experimentales asociados a determinar en un conjunto de pacientes si su columna vertebral es normal o anormal. Enseguida, se plantean las pruebas de análisis y resultados.

Palabras clave: algoritmos híbridos, redes neuronales artificiales, recocido simulado, predicción de datos, aprendizaje computacional.

Abstract

This paper is an advance of the research project entitled "Development of hybrid algorithms for data mining" and presents the use of a neural network with the simulated annealing algorithm to perform the prediction of a training data set. First, it addresses the problem to be solved, which is oriented to the analysis of the techniques defined for the hybrid algorithm. Then, the applied research methodology (descriptive-exploratory scientific with experimental approach) is justified. We performed a review of the techniques selected for the hybrid neuronal networks and simulated annealing technique which is applied to a set of experimental data associated with determining in a group of patients whether their spine is normal or abnormal. Then, the tests of analysis and results are presented.

Keywords: hybrid algorithms, artificial neural networks, simulated annealing, data prediction, machine learning.

[1] Magister en Ingeniería de Sistemas, Ingeniero de sistemas Docente Facultad Tecnológica Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. Correo electrónico: resalasr@udistrital.edu.co

[2] Magister en Ingeniería de Sistemas, Ingeniero de sistemas Docente Facultad Tecnológica Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. Correo electrónico: jerodriguezr@udistrital.edu.co

[3] Magister en Ciencias de la información y las comunicaciones, Ingeniera de sistemas Docente Facultad Tecnológica Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. Correo electrónico: clhernandez@gmail.com

1. Introducción

El volumen de datos que manejan las organizaciones cada vez es más grande, no sólo porque sus sistemas guardan transacciones más detalladas sino porque actualmente la información histórica también está siendo utilizada como soporte en la toma de decisiones. Para realizar tareas de proyección, las empresas se están valiendo, cada día con mayor frecuencia, de tecnologías de análisis de datos o minería de datos. Dentro de las opciones para aprovechar los datos que se recolectan en la operación diaria de las empresas, está la predicción de los datos. Existen diversas técnicas utilizadas para esta tarea, las cuales se usan dependiendo de algunos factores como la naturaleza de los datos. Así mismo, al verificarse el funcionamiento de las técnicas es posible usar una técnica híbrida que produzca mejoras a alguna de ellas y que al realizar la evaluación de indicadores se pueda establecer la pertinencia de su aplicabilidad y la viabilidad de su implementación.

Con este artículo se plantea el uso de la técnica híbrida de redes neuronales artificiales con recocido simulado, las cuales son usadas en tarea de predicción y optimización.

Para lograr los planteamientos establecidos se tuvo en cuenta las etapas de investigación [5] ya que el método aplicado en este proyecto fue la investigación científica con enfoque experimental.

El presente artículo se encuentra organizado en varias secciones, que recopilan el proceso que se llevó a cabo con este estudio. En primer lugar, se hace una descripción del problema a abordar. A continuación, se encuentra la metodología de investigación que se utilizó. Posteriormente, las técnicas analizadas junto con la técnica híbrida. Finalmente, se presenta el análisis de pruebas y resultados, y las conclusiones obtenidas.

2. Metodología

El método aplicado en este caso fue la investigación científica descriptiva-exploratoria con enfoque experimental. De acuerdo con el proceso formal de investigación, fue utilizado un método hipotético-deductivo en el cual se formuló una hipótesis, que a través de un razonamiento deductivo se validó de manera empírica. Se buscó establecer, con base en la experimentación, un mecanismo de ponderación de los indicadores de evaluación de las técnicas, de forma tal que fuera posible evaluar dicho mecanismo con el conjunto de datos.

Se realizaron las siguientes actividades, con el fin de utilizar la técnica híbrida basada en técnicas de optimización y predicción utilizando redes neuronales:

- Selección de las técnicas de para realizar la predicción y entrenamiento.
- Descripción de las técnicas para el caso de estudio y definición del algoritmo hibrido.
- Definición del conjunto de datos.
- Definición y aplicación de pruebas.
- Revisión de resultados de las pruebas.

Los instrumentos con los cuales se desarrolló esta investigación, son principalmente: casos de estudio y pruebas con datos experimentales.

2.1. Selección de las técnicas de predicción y entrenamiento

Las dos técnicas seleccionadas para el algoritmo híbrido son recocido simulado y redes neuronales. Con respecto a las ventajas se puede decir lo siguiente:

- Recocido simulado:
    - Simple en su implementación.
    - Basado en analogías de la física del templado de sólidos
    - Rápida convergencia a buenas soluciones
- Redes neuronales:
    - Son un aproximador universal.
    - Alto grado de no linealidad.
    - Alta tolerancia al ruido.
    - Habilidad de clasificar patrones para los cuales no ha sido entrenado

Considerando las desventajas es posible mencionar lo siguiente:

- Recocido simulado:

ISSN 1794-211X • e-ISSN 2322-939X • Vol 17, No 2 (julio-diciembre 2020). pp. 97-103. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

  - Requiere múltiples ajustes para que pueda lograr buenos resultados.
  - El resultado final depende de los valores iniciales de las variables a optimizar.
  - La convergencia en el óptimo global no está garantizada.
- Redes neuronales artificiales:
  - No funciona bien en problemas cuyos datos son categóricos.

Como se puede observar, las ventajas superan a las desventajas y con esta investigación se pretende comprobar que el recocido simulado es una técnica valida de ser utilizada para el ajuste de los pesos (entrenamiento) en las redes neuronales.

2.2. Descripción de las técnicas para el caso de estudio

Recocido simulado

El Simulated annealing o recocido simulado es una metaheurística que se basa en la analogía que puede existir entre un proceso de optimización combinatoria y un proceso termodinámico, conocido como recocido. Este proceso consiste en elevar la temperatura de un sólido cristalino con defectos hasta una temperatura determinada, que por lo general suele ser alta.
Posteriormente, se permite que el material se enfrié muy lentamente en un baño térmico. El proceso de enfriamiento viene descrito por una función de la temperatura, conocida como cola de enfriamiento, que generalmente suele ser continua y suave. Con este proceso se pretende que el sólido alcance una configuración de red cristalina lo más regular posible, eliminando durante este proceso los posibles defectos que tuviese originalmente. La nueva estructura cristalina se caracteriza por tener un estado de energía de la red mínimo [2].
Una de sus características más importantes es que se trata de un algoritmo de búsqueda local el cual se implementa para llegar a mejorar progresivamente las soluciones, Para ello se compara la solución anterior y se toma el primer movimiento que produce una mejora en la solución actual.
Otra característica que posee es utilizar una estructura de donde se estatiza el límite de soluciones anteriores para limitar el campo de acción de las soluciones siguientes, esto permite potenciar el algoritmo y permite encontrar soluciones en fronteras de menor tamaño.
Como la función objetivo está planteada desde el principio de la búsqueda, no presentará cambios durante el proceso, así el algoritmo no presentará fallos durante la toma de decisiones.
La cantidad de soluciones utilizadas por el recocido simulado para realizar la búsqueda pertenecen al tipo de metaheurísticas trayectoriales en donde el algoritmo parte de una solución inicial y se genera un camino o trayectoria en el espacio de búsqueda.

Redes neuronales artificiales

Una red neuronal artificial consiste de un conjunto simple de unidades de proceso, comunicadas para enviar señales a cada unidad a través de un alto número de conexiones [7].
El campo de las redes neuronales fue originalmente manejado por sicólogos y neurobiólogos quienes se dedicaron a desarrollar y evaluar el comportamiento de las neuronas. Más estrictamente hablando, una red neuronal es un conjunto de neuronas de entrada/salida conectadas entre sí, donde 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 [6].
Cuando se trabaja con redes neuronales artificiales, comúnmente se refiere a redes neuronales. El cerebro humano es altamente complejo, no lineal y procesa información en paralelo. Este tiene la capacidad para organizar el componente, conocido como neuronas, así como para realizar operaciones con certeza, tan rápido como las computadoras existentes hoy día. [4]

ISSN 1794-211X • e-ISSN 2322-939X • Vol 17, No 2 (julio-diciembre 2020). pp. 97-103. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

El algoritmo más popular para el ajuste de los pesos es el algoritmo de retropropagación que básicamente consiste en tomar el error de la salida en la red neuronales y en base a este error reajusta los pesos haciendo una propagación hacia atrás del mismo.

Técnica hibrida recocido simulado y redes neuronales

A continuación, se presenta la técnica hibrida propuesta de recocido simulado con redes neuronales:

Paso 1: Inicializar la temperatura (T) y generar de manera aleatoria los pesos de la red neuronal.
Paso 2: Calcular el error cuadrático medio de la configuración de pesos de la red neuronal para el conjunto de patrones de entrenamiento. Esta se tiene como solución actual del sistema.
Paso 3: Generar una solución vecina a la que se tiene como solución actual, es decir, generar un conjunto de pesos nuevos a la red neuronal vecinos a los que se tienen como solución actual.
Paso 4: Calcular el nuevo error cuadrático medio
Paso 5: Si Δ= (error nuevo – error actual) es menor o igual a cero, tomar la nueva solución como la solución actual al sistema.
Paso 5.1 Si Δ= (error nuevo – error actual) es mayor que cero, se calcula la expresión $e^{\Delta/T}$, si esta expresión es mayor que un número aleatorio que se genera entre cero y uno, se toma la nueva solución, como la solución actual al sistema, sino se rechaza.
Paso 6: Repetir los pasos 3, 4 y 5 un número determinado de iteraciones.
Paso 7: Decrecer la temperatura.
Paso 8: Repetir los pasos 3, 4, 5, 6 y 7 un número determinado de iteraciones.
Paso 9: Presentar el conjunto de pesos que se tenga como solución actual.

2.3. Definición del conjunto de datos

Los datos con los cuales se realizaron las pruebas finales de la estimación de datos usando la técnica híbrida de recocido simulado y redes neuronales, corresponden al conjunto de datos denominado "Vertebral Column Data Set" que incluye datos de 310 pacientes. Cada paciente está representado en el conjunto de datos por seis atributos biomecánicos derivados de la forma y orientación de la pelvis y la columna lumbar (en este orden): incidencia pélvica, inclinación pélvica, ángulo de lordosis lumbar, pendiente sacra, radio pélvico y grado de espondilolistesis (http://archive.ics.uci.edu/ml/datasets/vertebral+column). El atributo clase es normal o anormal.
La implementación de los algoritmos en referencia se realizó usando la herramienta MATLAB.

2.4. Definición y aplicación de pruebas

Preprocesamiento de los datos

La toma de decisiones se hace más eficiente si se define que los datos utilizados son de calidad y esto se logra si en la fase de preprocesamiento se detectan anomalías en los datos y corrigen a tiempo o se reduce el conjunto de datos para el análisis. [1] [3]
Para el entrenamiento de la red neuronal con recocido simulado se tuvo que normalizar todos los atributos del conjunto de datos, esta normalización se realizó haciendo uso de la ecuación (1)

$$V' = \frac{V+min}{max-min} * (0.9 - 0.1) + 0.1 \qquad (1)$$

Donde *V* es el valor de cada uno de los datos de cada atributo excepto el atributo clase, *min* es el valor menor de cada atributo, *max* es el valor mayor de cada atributo. El atributo clase toma dos posibles valores 0.1 para anormal y 0.9 para normal.
Luego se seleccionaron de manera aleatoria el 70% de los registros cuyo atributo clase fuera normal y 70% de los registros cuyo atributo clase fuera anormal, para el entrenamiento de la red neuronal y el 30% restante como datos de prueba.

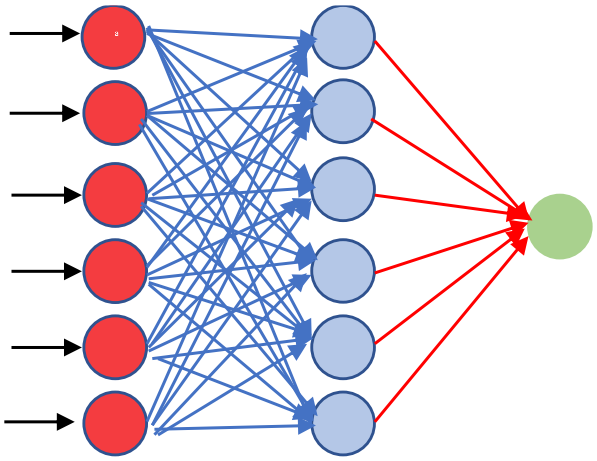
Configuración de la red neuronal

La red neuronal se configuró de la siguiente manera: seis neuronas en la capa de entrada, una neurona en la

ISSN 1794-211X • e-ISSN 2322-939X • Vol 17, No 2 (julio-diciembre 2020). pp. 97-103. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

capa de salida y seis de neuronas en la capa oculta, tal y como se muestra en la Figura 1.

Figura 1. Configuración de la red neuronal



Fuente: elaboración propia.

Configuración del recocido simulado

Para la ejecución del recocido simulado se utilizaron los siguientes parámetros:

- Temperatura inicial: 1000
- numero de iteraciones con la misma temperatura: 600
- numero de iteraciones con diferente temperatura: 230

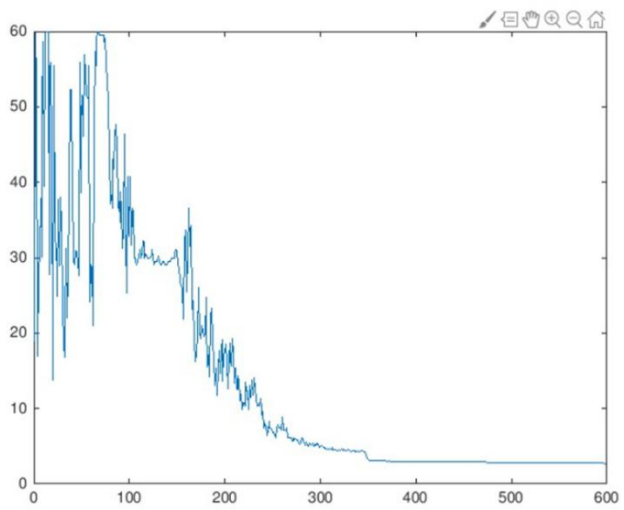
Esquema de reducción de temperatura: $T=\alpha T$ con $\alpha=0.97$.

3. Resultados

Para el entrenamiento, los datos se dividieron en dos: 70% para el entrenamiento y 30% de prueba, es decir, 217 registros de datos para el entrenamiento y 93 de prueba.

Luego de las configuraciones utilizadas tanto para la red neuronal como el algoritmo del recocido simulado y utilizando los datos de entrenamiento, se obtuvo la gráfica mostrada en la Figura 2 del error.

Figura 2. Salida del error de la red neuronal luego del entrenamiento.



Fuente: elaboración propia.

3.1. Matriz de confusión y parámetros de medición del entrenamiento

La matriz de confusión tiene la estructura presentada en la Tabla 1.

Tabla 1. Estructura de la matriz de confusión.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (VP)	Verdaderos Negativos (VN)

Fuente: elaboración propia.

La Tabla 2 presenta la matriz de confusión obtenida luego del entrenamiento de la red neuronal con el recocido simulado.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 17, No 2 (julio-diciembre 2020). pp. 97-103. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Tabla 2. Matriz de confusión para los datos de entrenamiento.

	PREDICCIÓN		
	NORMALES	ANORMALES	
NORMALES	54	16	70
ANORMALES	110	37	147
	164	53	

Fuente: elaboración propia.

Tal y como se puede observar en los resultados, los registros cuya salida es normal, en general los clasificó bien, cosa que no sucedió con los anormales.
A continuación, se presentan otros parámetros de medición del entrenamiento:

- Exactitud: La exactitud se calcula con la formula (2)

$$Exactitud = \frac{VP+VN}{Total} \quad (2)$$

Exactitud=0.4171

- Tasa de error: La Tasa de error se calcula con la formula (3)

$$Tasa\ de\ error = \frac{FP+FN}{Total} \quad (3)$$

Tasa de error=0.580

- Sensibilidad, Tasa de verdaderos positivos: La Sensibilidad se calcula con la formula (4)

$$Sensibilidad = \frac{VP}{Total\ Positivos} \quad (4)$$

Sensibilidad=0.771

- Especificidad, tasa de verdaderos negativos: La Especificidad se calcula con la formula (5)

$$Especificidad = \frac{VN}{Total\ Negativos} \quad (5)$$

Especificidad=0.228

- Precisión: La Precisión se calcula con la formula (6)

$$Precisión = \frac{VP}{Total\ Clasificados\ positivos} \quad (6)$$

Precisión=0.329

3.2. Matriz de confusión y parámetros de medición de los datos de prueba

La Tabla 3 presenta la matriz de confusión obtenida luego de calcular la salida de la red neuronal con los datos de prueba.

Tabla 3. Matriz de confusión para los datos de prueba.

	PREDICCIÓN		
	NORMALES	ANORMALES	
NORMALES	22	8	30
ANORMALES	43	20	63
	65	28	

Fuente: elaboración propia.

Tal y como se puede observar en los resultados, los registros cuya salida es normal, en general los clasificó bien, cosa que no sucedió con los anormales, tal y como sucedió con los datos de entrenamiento.
A continuación, se presentan otros parámetros de medición del entrenamiento:

- Exactitud=0.451
- Tasa de error=0.548
- Sensibilidad=0.617
- Especificidad=0.285

Precisión=0.338

ISSN 1794-211X • e-ISSN 2322-939X • Vol 17, No 2 (julio-diciembre 2020). pp. 97-103. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

4. Conclusiones

En el contexto de las técnicas de optimización y las redes neuronales para la predicción de datos, se seleccionaron las técnicas recocido simulado y red neuronal perceptrón multicapa. Con sus características, similitudes, debilidades y fortalezas, permitieron definir una técnica híbrida basada en los resultados obtenidos del análisis de las técnicas individuales y de los casos de estudio que confirmaron la posibilidad de utilizar una técnica híbrida de este estilo para la predicción de datos.

Los indicadores de Exactitud, tasa de error, sensibilidad, especificidad y precisión. Al aplicar dichos indicadores a los resultados de la técnica híbrida sobre los datos experimentales, permitieron establecer que no fueron buenos resultados para los datos anormales, si acertó con buena precisión los datos normales. Toca realizar más pruebas con diferentes combinaciones de red neuronal y algoritmo del recocido simulado para obtener mejor resultados.

Igualmente, se comprueba que la combinación de estas dos técnicas permite generar en promedio resultados aceptables para predicción de datos.

Referencias

[1] Clifton, C. (2006). Data Mining Course Overview. Disponible en http://www.cs.purdue.edu/homes/clifton/cs590d/Intro.ppt

[2] Duarte, M., Pantrigo F. y Gallego C. (2007). MetaHeurísticas. Madrid: Dykinson.

[3] Han, J. y Kamber, M. (2006). Data mining, Concepts and techniques. San Francisco, CA: Morgan Kaufmann. pp. 61 – 65, 110 – 127.

[4] Haykin, S. (1999). Neural Networks A comprensive foundation. USA

[5] Hernández, R., Fernández C. y Baptista, P. (2003). Metodología de la investigación. México: McGraw-Hill Interamericana.

[6] Kantardzic, M. (2001). Data Mining: concepts, models, methods, and algorithms. United States.

[7] Kröse, B. y Smagt, P (1996). An Introduction to Neural Network. Holland. p. 15.