San Marcos

MIEMBRO DE LA RED ILUMNO

REDES NEURONALES

REDES NEURONALES

REDES NEURONALES

Esta técnica de inteligencia artificial, en los últimos años, se ha convertido en uno de los instrumentos de uso frecuente para detectar categorías comunes en los datos, debido a que son capaces de detectar y aprender complejos patrones y características de los datos. Son ampliamente utilizadas en tareas relacionadas con el reconocimiento de patrones y sistemas de clasificación. Aunque son clasificadores muy precisos, su uso en minería de datos es aún área en estudio puesto que dan lugar a modelos de aprendizaje inestables. Estos modelos de redes neuronales son modelos matemáticos simples de interconexión entre neuronas artificiales. Las neuronas representan mediante simulación, los procesos que se dan sobre las neuronas del cerebro humano. Así, es entrenada a partir de un conjunto inicial de entrenamiento donde se generalizan patrones de predicción y clasificación. Cada neurona de la red procesa de forma independiente los datos que le llegan y reporta los resultados obtenidos del proceso interno a la siguiente capa de la red.

Para el pronóstico de series de tiempo, el modelo de predicción de orden p, tiene la forma general:

$$D_t = f(D_{t-l}, D_{t-l}, \ldots, D_{t-p}) + e_t$$

Las arquitecturas de redes neurales pueden ser entrenadas para predecir los valores futuros de las variables dependientes. Los requerimientos son el diseño del paradigma de la red y sus parámetros. El acercamiento de redes neurales de retroalimentación de capas múltiples consiste en una capa de entrada, una o varias capas escondidas y una capa de salida o resultado. Otro acercamiento es conocido como la red neural parcialmente recurrente, la cual puede aprender secuencias a medida que el tiempo transcurre y responde de manera diferente a los mismos patrones de estímulos de entrada a diferentes períodos de tiempo, dependiendo por supuesto de los distintos patrones de entrada. Ninguno de estos acercamientos es superior a cualquiera de los otros en cualquiera de los casos; sin embargo, una retroalimentación empapada que posea las características de una memoria dinámica, mejorará el funcionamiento de ambos acercamientos.

CONSIDERACIONES DE *OUTLIER*:

> Los *outliers* son algunas observaciones que no son bien ajustadas por el "mejor" modelo disponible. En la práctica, cualquier observación con residuos estandarizados con valor absoluto mayores a 2,5 es un candidato para ser considerado un *outlier*. En estos casos, se debería primero investigar el origen de los datos. Si no existe ninguna duda sobre la precisión o veracidad de las observaciones, entonces debería ser removido, y el modelo debería ser reajustado. (Alegret, 2008, párr. 220)

Siempre que los niveles de los datos sean considerados muy altos o muy bajos con respecto a los valores "usuales en el negocio", llamamos a estos valores *outliers*. Una razón matemática para ajustar estas ocurrencias es que la mayoría de las técnicas de pronóstico están basadas en promedios. Es bien sabido que las medias aritméticas son muy sensibles a los valores de los outliers; por lo tanto, algunas alteraciones en los datos deberían ser hechas antes de continuar. Una aproximación sería el reemplazar el *outlier* por el promedio de los dos niveles de ventas para los períodos, los cuales vienen inmediatamente antes y después del período en cuestión, y luego poner este número en el lugar del *outlier*. Esta idea es útil siempre que el *outlier* ocurre a la mitad o en una parte reciente de los datos. Sin embargo, si los *outliers* aparecen en la parte más antigua de los datos, se debería seguir una segunda alternativa, la cual es simplemente eliminar los datos e incluir los *outliers*.

En la ligereza de la relativa complejidad de algunas técnicas sofisticadas de pronóstico, nosotros recomendamos que la gerencia se dirija a través de una progresión evolucionaria para adoptar nuevas técnicas de pronóstico. Esto significa que es mejor que sea implementado un modelo de pronóstico simple bien entendido que a otro con todos los despliegues y presentaciones, pero que sea confuso en muchas facetas.

MODELAMIENTO Y SIMULACIÓN:

Los modelamientos y simulaciones dinámicas son la habilidad colectiva para entender el sistema y las implicaciones de sus cambios a través del tiempo, incluyendo el pronóstico. Los sistemas de simulación son una mímica de la operación del sistema real, tal como las operaciones diarias de un banco, o el valor de una determinada acción en la bolsa de valores durante un periodo de tiempo específico. Mediante las corridas de simulación para avanzar en decisiones futuras, los gerentes pueden encontrar fácilmente como el sistema podría comportarse en el futuro, por lo tanto, las decisiones podrían ser juzgadas como apropiadas.

LOS MODELAMIENTOS Y SIMULACIONES DINÁMICAS SON LA HABILIDAD COLECTIVA PARA ENTENDER EL SISTEMA Y LAS IMPLICACIONES DE SUS CAMBIOS A TRAVÉS DEL TIEMPO, INCLUYENDO EL PRONÓSTICO.

En el campo de las simulaciones, el concepto del "principio de la equivalencia computacional" tiene implicaciones favorables para los tomadores de decisiones. Las experimentaciones simuladas aceleran y reemplazan efectivamente la ansiedad de "esperar para ver qué sucede" descubriendo nuevas formas y explicaciones para comportamientos futuros del sistema real.

MODELOS PROBABILÍSTICOS:

> El uso de técnicas probabilísticas, tales como los métodos de investigación de mercadeo, para lidiar con incertidumbre, ofrece un rango de resultados probables para cada grupo de eventos. Por ejemplo, se podría desear identificar los prospectos compradores de un nuevo producto dentro de una comunidad de tamaño N. Del resultado de una encuesta, se podría estimar la probabilidad de vender p, y luego estimar el tamaño de las ventas totales Np con un cierto nivel de confianza. (Alegret, 2008, párr. 220)

UNA APLICACIÓN:

"Suponga que deseamos pronosticar las ventas de una nueva pasta de dientes en una comunidad de 50.000 amas de casas. Una muestra gratis es suministrada a 3.000 de ellas que fueron seleccionadas de manera aleatoria, y luego 1.800 de ellas indicaron que comprarían el producto." (Alegret, 2008, párr. 220)

Ejemplo de la aplicación del modelo

La aplicación de redes neuronales a la predicción de series temporales ha atraído la atención de mucha gente relacionada con los mercados financieros de todo el mundo. (Garrido, 2012, p. 65)

Sin embargo, el hecho de que la evolución de los mercados dependa de multitud de variables, muchas de las cuales son difíciles de cuantificar, hace que sea complicado encontrar situaciones en las que únicamente un análisis numérico de los datos históricos permita realizar buenas predicciones.

Como aplicación práctica de este tipo de redes neuronales, consideraremos el análisis de una acción en la bolsa noruega. Seguiremos paso a paso la metodología básica de la aplicación de esta técnica.

1. Definición del problema

Deseamos conocer la evolución de la acción PGS a dos días. El primer paso requiere decidir qué variables serán alimentadas a la red neuronal. Hemos optado por considerar como variables de entrada: la propia serie PGS, el índice Nasdaq, la cotización de la misma empresa en el mercado americano (PGSUS), el índice de la bolsa noruega y los tipos de interés en Estados Unidos.

2. Reprocesamiento de las variables

La red neuronal podrá extraer información útil si las variables que la alimentan están correctamente reprocesadas. Hemos decidido considerar transformaciones de las variables del tipo considerado en el chartismo (medias móviles, ROC, etc.) y en otras técnicas de análisis de datos (Gumbel, etc.) (Alegret, 2008, párr. 221)

3. Entrenamiento de las redes

Las redes neuronales son entrenadas evitando sobre entrenamiento y falta de generalización. Es conveniente considerar un entrenamiento que corresponda a mostrar cada patrón unos pocos miles de veces a la red neuronal. El entrenamiento breve no logra captar la ley que subyace al mercado. Un entrenamiento demasiado largo hace que la red aprenda los errores (no la ley general) de cada patrón. Una forma de evitar sobre entrenamiento es construir modelos alternativos con redes neuronales pequeñas. Una red pequeña no tiene grados de libertad suficientes para llegar a estar sobreentrenada.

4. Evaluación del modelo

Una vez que nuestra red neuronal es entrenada disponemos de un modelo de predicción. Podemos a continuación rastrear el éxito o fracaso de este modelo en los últimos meses. Es importante lograr que nuestro modelo además de ser rentable sea robusto frente pequeños cambios de variables o de estructura de la red neuronal. El modelo debe también describir con igual éxito tanto las subidas como las bajadas del mercado. Por último, el modelo debe proporcionar resultados similares a lo largo del tiempo. Si cualquiera de estos requisitos no se cumple podemos fácilmente haber construido un modelo que tendrá pobres resultados.

5. Definición de estrategia de actuación

El modelo neuronal debe considerarse como una herramienta de predicción que se complementa con otros métodos alternativos. Es muy posible que no todos los datos del mercado queden reflejados en las pocas series que alimentan a la red. Una correcta estrategia de actuación debería contar con los condicionantes de la agresividad del inversor, de su liquidez, etc.

EL MODELO NEURONAL DEBE CONSIDERARSE COMO UNA HERRAMIENTA DE PREDICCIÓN QUE SE COMPLEMENTA CON OTROS MÉTODOS ALTERNATIVOS. ES MUY POSIBLE QUE NO TODOS LOS DATOS DEL MERCADO QUEDEN REFLEJADOS EN LAS POCAS SERIES QUE ALIMENTAN A LA RED.

Los resultados obtenidos por las redes neuronales entrenadas siguiendo los pasos anteriores ofrecen resultados muy satisfactorios. Sobre un período de seis meses se logran beneficios de un 20%, ignorando comisiones de compra. Los beneficios se obtienen tanto en operaciones de compra como de venta (en el mercado de futuros). (Alegret, 2008, párr. 223).

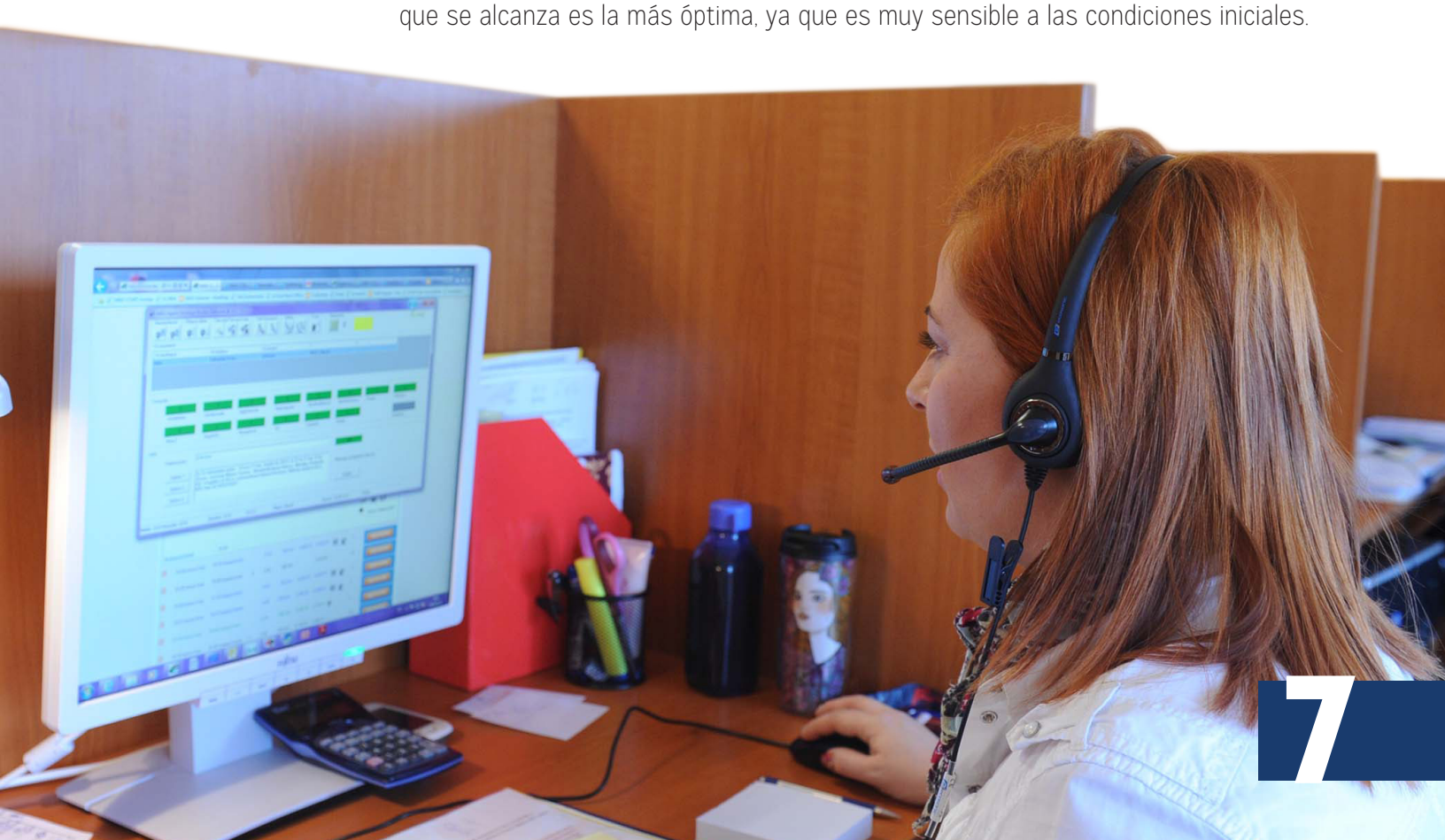
ANÁLISIS DE MODELOS DE *CLUSTER*, LÓGICA DIFUSA *K-MEANS*

El algoritmo de las K-medias es uno de los algoritmos de aprendizaje no supervisado más simples para resolver el problema de la clusterización. El procedimiento aproxima por etapas sucesivas un cierto número (prefijado) de clusters haciendo uso de los centroides de los puntos que deben representar.

El algoritmo se compone de los siguientes pasos:

- Sitúa *K* puntos en el espacio en el que "viven" los objetos que se quieren clasificar. Estos puntos representan los centroides iniciales de los grupos.
- Asigna cada objeto al grupo que tiene el centroide más cercano.
- Tras haber asignado todos los objetos, recalcula las posiciones de los *K* centroides.
- Repite los pasos 2 y 3 hasta que los centroides se mantengan estables. Esto produce una clasificación de los objetos en grupos que permite dar una métrica entre ellos (Sancho, 2015, párr. 2).

Aunque se puede probar que este algoritmo siempre termina, no siempre la distribución que se alcanza es la más óptima, ya que es muy sensible a las condiciones iniciales.



FILTRO DE KALMAN

El filtro de Kalman es un algoritmo para la actualización secuencial de una proyección lineal en un sistema dinámico, el cual está representado en una fase espacial. Las aplicaciones del filtro de Kalman son las de transformar el sistema de una representación de las dos fórmulas siguientes a una forma más razonable:

EL FILTRO DE KALMAN ES UN ALGORITMO PARA LA ACTUALIZACIÓN SECUENCIAL DE UNA PROYECCIÓN LINEAL EN UN SISTEMA DINÁMICO, EL CUAL ESTÁ REPRESENTADO EN UNA FASE ESPACIAL.

Sea $x_{t+1}=Ax_t+Cw_{t+1}$, y $y_t=Gx_t+v_t$ en el cual: A, C, y G son matrices conocidas como funciones del parámetro q sobre el cual la inferencia es deseada donde: t es un número entero, usualmente el tiempo indexado; x_t es una variable de estado verdadero, escondido de los econometristas; y_t es una medición de x con un factor escalado G, y los errores de medición y_t, w_t son innovaciones a los procesos escondidos x_t, $E(w_{t+1}w_t')=1$ por normalización (donde, ' significa la transpuesta), $E(v_tv_t)=R$, una matriz desconocida, una estimación la cual es necesaria pero es auxiliar al problema de interés, el cual es el obtener una estimación de q. El filtro de Kalman define dos matrices S_t y K_t de forma tal que el sistema descrito anteriormente puede ser transformado en el siguiente, en el cual las estimaciones e inferencias sobre q y R son más directas, es decir, mediante el análisis de regresión:

Sea $z_{t+1}=Az_t+Ka_t$, y $y_t=Gz_t+a_t$ de donde z_t está definida a ser $E_{t-1}x_t$, a_t está definida a ser $y_t-E(y_{t-1}y_t$, y K está definida a ser el límite K_t cuando t se acerca al infinito.

La definición de estas dos matrices S_t y K_t es en sí misma la definición de los filtros de Kalman: $K_t=AS_tG'(GS_tG'+R)^{-1}$, y $S_{t-1}=(A-K_tG)S_t(A-K_tG)'+CC'+K_tRK_t'$, K_t es comúnmente llamada la ganancia de Kalman. (Barcellos, 2006, párr. 34)

Este método se puede dividir en:

1. Algoritmos genéticos

Los algoritmos genéticos imitan la evolución de las especies mediante la mutación, reproducción y selección, como también proporcionan programas y optimizaciones que pueden ser usadas en la construcción y entrenamiento de otras estructuras como es el caso de las redes neuronales. Además, los algoritmos genéticos son inspirados en el principio de la supervivencia de los más aptos.

2. *Clustering* (Agrupamiento)

Agrupan datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud, de manera que las clases sean similares entre sí y distintas con las otras clases. Su utilización ha proporcionado significativos resultados en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de sistemas. Este método debido a su naturaleza flexible se puede combinar fácilmente con otro tipo de técnica de minería de datos, dando como resultado un sistema híbrido. (Moreno, 2007, párr.6).

Un problema relacionado con el análisis de cluster es la selección de factores en tareas de clasificación, debido a que no todas las variables tienen la misma importancia a la hora de agrupar los objetos. Otro problema de gran importancia y que actualmente despierta un gran interés es la fusión de conocimiento, ya que existen múltiples fuentes de información sobre un mismo tema, los cuales no utilizan una categorización homogénea de los objetos. Para poder solucionar estos inconvenientes es necesario fusionar la información a la hora de recopilar, comparar o resumir los datos.

3. Aprendizaje automático

Esta técnica de inteligencia artificial es utilizada para inferir conocimiento del resultado de la aplicación de alguna de las otras técnicas antes mencionadas.

MODELOS DE ASOCIACIÓN DE VARIABLES

En la investigación de minería de datos nos encontramos con frecuencia con datos o variables de tipo cualitativo, mediante las cuales un grupo de individuos se clasifican en dos o más categorías mutuamente excluyentes. Las proporciones son una forma habitual de expresar frecuencias cuando la variable objeto de estudio tiene dos posibles respuestas, como presentar o no un evento de interés (enfermedad, muerte, curación, etc.). Cuando lo que se pretende es comparar dos o más grupos de sujetos con respecto a una variable categórica, los resultados se suelen presentar a modo de tablas de doble entrada que reciben el nombre de tablas de contingencia. Así, la situación más simple de comparación entre dos variables cualitativas es aquella en la que ambas tienen solo dos posibles opciones de respuesta (es decir, variables dicotómicas). (Barcellos, 2006, párr. 23)

Las tablas de contingencia son una de las herramientas más antiguas y conocidas de la estadística, por lo que su utilización rutinaria puede llevar aparejada una cierta despreocupación, que es contraria al cuidado y meticulosidad con el que siempre deben analizarse los datos, sin abandonarnos a la tarea simple de introducir datos en un programa informático y limitarnos a transcribir mecánicamente los resultados obtenidos, sin mayor análisis, restringiendo además nuestra mirada a los resultados con los que estamos familiarizados, y olvidándonos del resto de información que quizás no entendemos.

Las pruebas de significación del chi^2 permiten contrastar si es razonable pensar que la relación observada entre las variables puede ser simplemente atribuida al azar. En el nivel de significación influye, como en cualquier otra prueba estadística, no solo la importancia o magnitud de la asociación, sino también el tamaño de la muestra y en ocasiones otros parámetros. Es posible obtener un resultado estadísticamente significativo con una débil asociación, si el tamaño de muestra es suficientemente grande, y viceversa, si la muestra es pequeña una asociación importante puede no llegar a ser estadísticamente significativa.

Esto es algo que es de dominio común, y es universalmente aceptado en cualquier otra prueba estadística que nunca se debe presentar únicamente un valor de P, sino que este debe acompañar a algún parámetro que exprese la magnitud del resultado, o mejor aún un intervalo de confianza para el efecto observado. Sin embargo esto, que es práctica habitual en el resto de pruebas estadísticas, no se lleva a cabo con las pruebas de asociación en tablas de contingencia, salvo que éstas sean 2x2, en cuyo caso se suele presentar como medida de la asociación alguna medida relativa como el odds ratio o el riesgo relativo, o bien una diferencia de proporciones. (Barcellos, 2006, párr. 34)

EXTRACCIÓN DE REGLAS

En ámbitos como medicina, ciencia, ingeniería o marketing se acumula cada vez una mayor cantidad de datos, clave para nuevos e importantes descubrimientos. Por ejemplo, en Biología molecular se espera utilizar la gran cantidad de información que se está tratando de reunir actualmente para comprender mejor la estructura y la función de los genes. En el pasado, métodos tradicionales de biología molecular permitían a los científicos el estudio de unos pocos genes al mismo tiempo en un experimento concreto, mientras que hoy en día, gracias al desarrollo de las técnicas basadas en microarrays, es posible comparar el comportamiento de miles de genes en diversas situaciones. Estas comparaciones pueden ayudar a determinar la función de cada gen y, quizás, determinar qué genes causan ciertas enfermedades. Sin embargo, la presencia de ruido y la gran dimensionalidad de los datos hacen necesario el desarrollo de nuevos tipos de análisis. (Sistemas Adaptativos y Bioinspirados en Inteligencia Artificial, 2016, p. 47).

Data Mining o minería de datos consiste en descubrir automáticamente información útil en grandes repositorios de datos. Este tipo de técnicas se utilizan, por lo tanto, para analizar grandes bases de datos en busca de nuevos patrones que sean útiles y que, de otro modo, no serían descubiertos. Además, permiten predecir la salida de una observación futura, como, por ejemplo, si un nuevo cliente gastará más o menos de una cierta cantidad de dinero.

Las técnicas tradicionales de análisis de datos suelen encontrarse con ciertos problemas a la hora de tratar de superar nuevos retos que ofrecen nuevos conjuntos de datos. Algunos retos específicos que motivaron el desarrollo de la minería de datos son los siguientes:

1. **ESCALABILIDAD**
2. **ALTA DIMENSIONALIDAD**
3. **DATOS HETEROGÉNEOS Y COMPLEJOS**
4. **LA PROPIEDAD Y DISTRIBUCIÓN DE LOS DATOS**
5. **ANÁLISIS NO TRADICIONAL**

Las tareas de *Data Mining* se suelen dividir en dos grandes categorías:

- **Tareas predictivas**, cuyo objetivo es predecir el valor de un atributo (característica) en particular, basándose en los valores de otros atributos.
- **Tareas descriptivas**, cuyo objetivo es obtener patrones que representen las relaciones subyacentes existentes en los datos.

Estas tareas tienen especial relevancia en el ámbito biomédico, ya que ayudarían tanto en el diagnóstico y prevención de enfermedades, como en un mejor conocimiento de sus características. A través de tareas predictivas se podría realizar diagnóstico y prevención y mediante tareas descriptivas se podrían obtener las características del ADN que podrían predisponer a un paciente a padecer una enfermedad. Esto último se podría llevar a cabo mediante la extracción de reglas de asociación a partir de datos genéticos. Este tipo de reglas asocian ciertas características presentes en el ADN de un paciente con, por ejemplo, el desarrollo de una enfermedad o el ser afectado por un efecto secundario de un medicamento. Por ello, se presenta una aplicación capaz de obtener reglas de asociación a partir de datos genéticos y capaces de predecir el estado de un paciente, sano o enfermo en relación con una enfermedad.

Para obtener dichas reglas, la aplicación analizará el conjunto de datos proporcionados como entrada utilizando una técnica de Computación Evolutiva: los Algoritmos Genéticos (AAGG). Se trata de una técnica de inteligencia artificial basada en la Teoría de la Evolución de Charles Darwin, de tal forma que se inspira en la evolución biológica y su base genético-molecular. Este tipo de algoritmos hacen evolucionar una población a través de acciones aleatorias similares a las que existen en la evolución biológica, como la mutación y la recombinación genética, y a través de un mecanismo similar a la selección natural. (Barcellos, 2006, párr. 34)

La aplicación desarrollada, además de extraer reglas, puede ser utilizada para realizar clasificaciones. Las características (o atributos) proporcionadas serán, pues, analizadas para buscar relaciones entre ellas y con base en estas relaciones se realizará la clasificación de los datos de entrada, cubriendo los dos grandes categorías de tareas de minería de datos. Esto abre un gran abanico de posibilidades de aplicación en ámbitos completamente diferentes y con diversos objetivos, permitiendo tanto la predicción como la obtención de estructuras relevantes de gran cantidad de datos que pueden poseer ruido o, incluso, en los que puede faltar cierta información. (Barcellos, 2006, párr. 34)

REFERENCIAS BIBLIOGRÁFICAS

Alegret, M., Herrera, M. y Grau, R. (2008). Las técnicas de estadística espacial en la investigación salubrista. Revista Cubana Sal Públ [revista en la Internet]. Recuperado de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-34662008000400003&lng=es

Barcellos, C. y Buzai, G. (2006). La dimensión espacial de las desigualdades sociales en salud: aspectos de su evolución conceptual y metodológica. México: Universidad Nacional de Luján: Anuario de la División Geografía.

Benzecri, J. (2015). Gestión del conocimiento y minería de datos. París: Dunod.

Berry, M. y Linoff, G. (2014). Data Mining Techniques for Marketing Sales and Customer Support. USA: John Wile& Sons

Chatfield, C. y Collins, A.J. (1999). Introduction to multivariate analysis. London: Chapman and Hall.

Kamber M. (2006). Data mining: concepts and techniques. Morgan Kaufmann;

Hernández R. (2004). USA: Centricity Solution for Marketing.

Hoaglin D.C., Mosteller F., Tukey J.W. (2005). N.Y: Exploring Data tables. Trends and Shapes, Wiley.

Kulldorff, M. y Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. Customer Support. USA: John Wile& Sons

Ester, M., Kriegel, H., y Sander, J (1999). Knowledge discovery in spatial databases. KI-99. Advanc Artif Intellig.

Jambu, M. (2000). Classification Automatique pour l'Analyse des données. París: Dunod.

Johnson Dallas, E. (2013). Métodos multivariados aplicados al análisis de datos. México: Thomson editores.

Johnson, R.A. y Wichern Dean, W. (2010). Applied Multivariare Statistical Analysis. 3rd De. USA: Prentice Hall Inc.

Lebart, L., Morineau, A. y Tabard, N. (2000). Téchniques de la description statistique. París: Dunod.

Lebart, L., Morineau, A. y Piron, M. (2005). Statistique exploratoire multidimensionelle. París: Dunod.

Macoe, J. (2016). 5 de los mejores software de minería de datos de Código Libre y Abierto. Recuperado de: Http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/

Monografía Nº 27 Serie de matemática. USA: O.E.A.Washington.

Tecnologías-información.com. (2015). Minería de Datos. Recuperado de: http://www.tecnologias-informacion.com/mineria-de-datos.html

Tecnologías-información.com. (2015). Minería de Datos. Recuperado de: http://www.tecnologias-informacion.com/mineria-de-datos.html

Vinnakota, S. y Lam, N. (2006). Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. Internat J Heal Geogr.

Zhao, F., Zhu, R., Zhang, L., Zhang, Z, Li, Y. y He. M. (2011). Application of satscan in detection of schistosomiasis clusters in marshland and lake areas. Zhongguo xue xi chong bing fang zhi za zhi. Alemania: Chin J Schistosom Contr.