UNIVERSIDAD AUTONOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FISICO MATEMATICAS

Alan Zamarron Medrano 1625654

Minería de Datos

Técnicas de Minería de Datos

Clustering

Este método es una técnica dentro de la disciplina de la Inteligencia Artificial, la cual identifica de manera automática agrupación de acuerdo a una medida de similitud entre ellos. Su objetivo fundamental es identificar clústeres de elementos, tales que: su similitud media entre elementos del mismo clúster sea alta (conocido como Similitud intra-cluster alta) y su similitud media entre elementos de distintos clústeres sea baja (conocido como Similitud inter-cluster baja).

Algunas de las aplicaciones del análisis de clustering son: marketing; ayuda a los especialistas en marketing a encontrar grupos distintivos entre sus clientes, y así mejorar sus programas de marketing específicos, biología; ayuda a definir clasificaciones de plantas y animales o a identificar genes con funcionalidades similares, Web; útil para clasificar documentos en la web para el descubrimiento de información, Detección de fraudes; útil en aplicaciones de detección de outliers, como la detección de fraudes de tarjetas de crédito.

Las características importantes del clustering son: Deben poder manejar conjuntos de datos pequeños, así como también grandes sin problemas, capacidad para manejar diferentes atributos, debe ser independiente del orden de entrada de los datos y finalmente los resultados obtenidos de los clusters deben ser completamente interpretables, lógicos y utilizables.

Predicción

El análisis predictivo es un área de la minería de datos que consiste en la extracción de información y su utilización para predecir tendencias y patrones de comportamiento, pudiendo aplicarse sobre cualquier evento desconocido, sea cual sea el tiempo en el que se esté. Para llevar a cabo el análisis predictivo es indispensable disponer de una considerable cantidad de datos, tanto actuales como pasados, para poder establecer patrones de comportamiento y así inducir conocimiento. Los datos son la fuente de la que se obtienen las variables, las relaciones entre ellas, el conocimiento inducido o los patrones de comportamiento identificados, convirtiéndose en un elemento vital de todo análisis predictivo.

El modelo predictivo se podrá utilizar para predecir que probabilidades hay de que una persona reaccione de una manera determinada. Una vez introducidos los datos de la persona y se aplique el modelo predictivo se obtendrá una calificación que indicara la probabilidad de que se produzca la situación estudiada por el modelo.

Técnicas aplicables al análisis predictivo: Técnicas de regresión (Regresión Lineal, Arboles de clasificación y Regresión, Curvas de Regresión adaptativa multivariable) y Técnicas de Aprendizaje Computacional (Redes Neuronales, Maquinas de vectores de soporte, Naïve Bayes y K-means)

Outliers

El análisis de la calidad de los datos es de gran importancia para las organizaciones, ya que datos con problemas pueden conducir a decisiones erróneas con consecuencias como pérdida de dinero, tiempo y credibilidad. Entre los posibles problemas que pueden presentar los datos, se encuentran los conocidos como Outliers. Un Outlier es una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente.

Hay diferentes tipos de Outliers, entre esos están: Casos atípicos que surgen de un error de procedimiento, tales como entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de datos, en caso contrario, deberían eliminarse del análisis o recodificarse como datos ausentes; Observaciones que ocurren como consecuencia de un acontecimiento extraordinario, el outlier no representa ningún segmento valido de la población y puede ser eliminado del análisis; Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables, estas observaciones deberían ser retenidas en el análisis pero estudiando que influencia ejercen en los procesos de estimación de los modelos considerados; datos extraordinarios para las que el investigador no tiene explicación, en estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados.

Patrón Secuencial

Consiste en describir patrones interesantes, útiles e inesperados en una base de datos, la tarea de minería de patrones secuenciales se especializa en analizar datos secuenciales y descubrir patrones secuenciales, esto se puede traducir a encontrar subsecuencias de un conjunto de secuencias.

La minería de patrones periódicos es una tarea importante ya que de manera periódica pueden aparecer diferentes tipos de datos, y conviene entenderlos para tomar decisiones estratégicas.

Análisis Biológico Secuencial, este compara y analiza las secuencias biológicas, lo cual resulta crucial en el análisis bio informático, estas secuencias pueden estar formadas de nucleoides o amino ácidos.

Otro ejemplo de la aplicación de patrones secuenciales puede presentarse en análisis de textos. Un conjunto de frases de un texto puede ser vistas como la base de datos de secuencias, y el objetivo de patrón secuencial es encontrar las palabras más utilizadas en el texto.

Reglas de Asociación

Las Reglas de Asociación cuentan con algoritmos de asociación, estos tienen como objetivo encontrar relaciones dentro de un conjunto de transacciones ítems o atributos que tienden a ocurrir de forma conjunta. A cada uno de los eventos o elementos que forman parte de una transacción se le conoce como ítem y aun conjunto de ellos se les conoce como itemset, una transacción puede estar formada por uno o varios ítems, en el caso de ser varios, cada posible subconjunto de ellos es un itemset distinto.

Una base de datos transaccional se puede representar con las sig. métricas: una lista; representa cada transacción como una fila, cada fila lista los artículos comprados por el consumidor y es una transacción por lo que cada fila puede tener un numero diferente de columnas, una representación vertical; es la forma mas eficiente de guardar los datos de tamaño mas industrial o comercial, este ocupa solo 2 columnas y la representación Horizontal; se representa como una matriz binaria, cada fila de una matriz representa una transacción, y cada columna representa un artículo, si un articulo esta presente se representa como 1, en caso contrario se representa con 0.

El algoritmo mas importante de las reglas de asociación es el A priori el cual sus objetivos son: Identificar todos los itemsets que ocurren sobre un determinado limite y convertir esos itemsets frecuentes en reglas de asociación.

Regresión Lineal

En la regresión lineal buscamos una variable aleatoria simple como x, y, n, etc, en teoría el valor de esta variable aleatoria esta influenciado por los valores tomados por una o más variables, la variable a buscar se denomina como: variable dependiente (o Respuesta), las variables que influirán en el resultado son variables independientes, predictoras o regresoras.

Al realizar una predicción los regresores no se tratan como variables al azar, son entidades que asumen valores diferentes, pero no al azar.

Para realizar la regresión lineal se necesita realizar el modelo lineal, con los datos proporcionados en el problema que se indique en el momento, pero también se tiene que conocer el valor del error, o posible error, y realizar una predicción de los datos.

Otro termino a conocer es el error estándar residual, que este es la desviación estándar del termino del error (desviación de la parte de datos ya que el modelo no es capaz de explicar por falta de información o mas datos que este adicionados a dicho problema que se quiera investigar)