UNIVERSIDAD AUTONOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FISICO MATEMATICAS

Luis Rodolfo Torres Contreras 1742143

Minería de Datos

Técnicas de Minería de Datos

Contenido

Clustering

Análisis de grupos o agrupamiento es la tarea de agrupar objetos por similitud, en grupos o conjuntos de manera que los miembros del mismo grupo tengan características similares

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

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.

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 del 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

Un outlier sería una observación dentro de una muestra o una serie temporal de datos que no es consistente con el resto. Imaginemos por ejemplo que estamos midiendo la altura de los alumnos de una clase.

En situaciones como esta en la que existen valores anormales y que se distancian sustancialmente del resto, la mediana es una mejor estimación para saber en qué punto se concentran un mayor número de observaciones.
En la mayoría de los casos, los outliers tienen influencia en la media , pero no en la mediana , o la moda . Por lo tanto, los outliers son importantes en su efecto en la media.

No hay una regla para identificar a los outliers. Pero algunos libros refieren un valor como un outlier si este es mayor que 1.5 veces el valor del rango intercuartil más allá de los cuartiles .

También graficando los datos en una recta numérica como una gráfica de puntos, nos ayuda a identificar a los outliers.

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

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, 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.

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)