Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Minería de Datos

Prof. Mayra Cristina Berrones Reyes

“Técnicas de minería de datos”

Sergio Velázquez Rivera

1805244

Gpo: 003

2 de Octubre del 2020, San Nicolas de los Garza

REGLAS DE ASOCIACIÓN

Es una técnica que se utiliza en la inteligencia artificial (data mining) que es el proceso descubrimiento de tendencias o patrones en grandes bases de datos y lo que hace una regla de asociación en general es describir una regla de asociación entre los elementos de un conjunto de datos relevantes dicho de una manera más formal es la búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de elementos u objetos en bases de datos de transacciones, bases de datos relacionales y otros repositorios. Se puede aplicar para el análisis de datos de la banca, para el cross-marketing y el diseño de catálogos.

Para obtener las reglas de asociación es importante destacar que la confianza no tiene una propiedad anti monótona, además que para cada ítem se obtendrán los posibles sub-sets, de estos se creará la regla para después descartar aquellos que no superen la regla de mínimo de confianza.

Para el enfoque de fuerza bruta, es que teniendo listas todas las reglas de asociación, comprobando el soporte y la confianza, se eliminen las reglas que fallan según los umbrales.

Las estrategias de generación de los elementos frecuentes que aparecen con mayor frecuencia se utiliza el Principio Priori, el cual, reduce el número de candidatos (si es frecuente entonces todos sus subconjuntos también serán frecuentes). Este algoritmo fue uno de los primeros en ser desarrollados y actualmente es uno de los más empleados, se compone de 2 etapas. El primero es Identificar los ítems sets que ocurren con mayor frecuencia y el segundo convertir esos ítems sets frecuentes en reglas de asociación.

Otro método para la generación de los elementos frecuentes es la Class transformation, esta consiste en cómo se escanean y analizan los datos.

DETECCIÓN DE OUTLIERS

La detección de outliers o valores atípicos, pertenece a la categoría descriptiva que revisa la minería de datos donde su objetivo es encontrar patrones que de o maque un resumen de las relaciones ocultas de los datos estudiando el comportamiento de valores externos que difieren del patrón general de una muestra.
Los Valores atípicos son valores diferentes a las observaciones del mismo grupo de datos.
Los datos atípicos ocasionados por errores de entrada y procedimiento, acontecimientos extraordinarios, valores extremos o por causas no conocidas.

Existen distintos tipos de técnicas para detectarlos y se pueden dividir en dos categorías principales las cuales son métodos univariantes de detección y métodos multivariables. Las técnicas de detección para los valores atípicos son la prueba de GRUBBS, prueba DIXION , prueba de TUKEY, análisis de Valores y regresión Simple.

En la minería de datos se puede aplicar para la detección de fraudes financieros, la tecnología informática y telecomunicaciones, nutrición y salud, negocios, entre otros.
Los Outliers, pueden significar error, por ejemplo, si tenemos un grupo de “edades de personas” y tenemos una persona con 160 años, seguramente sea un error de carga de datos. Puede significar límites, que son valores que se escapan del “grupo medio”, pero queremos mantener el dato modificado, para que no perjudique al aprendizaje del modelo. Puede significar punto de interés, que podrían ser los casos anómalos los que queremos detectar y que sean nuestro objetivo.

REGRESIÓN LINEAL

La primera forma de represión lineal documentada fue el método de los mínimos cuadrados formalmente se utilizó para revisar las estaturas y cómo influía la estatura de padres con la estatura de los hijos (ya que si había padres de estatura baja, sus hijos tendían a crecer y volver a regresar a la media de estatura). Concretamente una regresión es un modelo matemático para determinar el grado de dependencia entre una o más variables, es decir, si existe relación entre ellas. Se pueden presentar dos tipos de regresión lineal, primero la regresión lineal que nos dice como una variable influye a otra y la regresión lineal múltiple que nos presenta diversas variables influyen a otra.

Hablando en minería de datos, las regresiones lineales se encuentran dentro de la categoría predictivo, esta categoría tiene como objetivo analizar los datos de un conjunto y en base a eso nos ayuda para poder predecir el mejoramiento de nuestras decisiones. Permite examinar la relación entre dos o más variables e identificar cuáles son las que tienen mayor impacto en un tema de interés. La variable independiente es el factor más importante y la variable independiente es el factor que uno cree que puede impactar en nuestra variable dependiente.

La idea de la regresión lineal consiste en obtener una ecuación de la forma $y=mx+b$ que se ajuste más a los datos, donde m es la pendiente de los datos y $b=\overline{y}-m\overline{x}$.
Para determinar qué tan bueno es el ajuste, existen diferentes parámetros estadísticos, pero en este caso se utiliza el coeficiente de determinación $R=(\sigma xy)/(\sigma x\ \sigma y)$.
Se necesita saber si esta regresión es significativa para tener idea si existe estas relaciones entre cada uno. Para saber si lo es, se usa la prueba de significancia y que la R2 ajustada sea grande para tener una mejor aproximación al modelo.

CLUSTERING

El Clustering o Agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares. Son las que utilizando algoritmos matemáticos se encargan de agrupar objetos, usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente y a la vez diferente entre los miembros de las diferentes clases. Esta técnica es la más utilizada en algoritmos matemáticos se encargan de agrupar objetos.

Existen muchas aplicaciones para esta técnica las cuales entran el estudios de terremotos: los epicentros del terremoto observado deben agruparse a lo largo de fallas continentes, las planificaciones de las ciudades: identificación de grupos de casas según su tipo de casa, valor y ubicación geográfica, de marketing: ayuda a los profesionales de marketing a descubrir distintos grupos en sus bases de clientes, aseguradoras: identificación de grupo de aseguradoras de seguros de automóviles en un alto costo promedio de reclamo, uso del suelo: identificación de áreas de uso similar de la tierra en una base de datos de observación de la tierra, etc.

Los métodos de agrupación se clasifican por asignación jerárquica frente a punto, datos numéricos y/o simbólicos, determinista vs probabilística, exclusivo vs superpuesto, jerárquico vs plano, de arriba abajo y viceversa. Entre estos se encuentra: Asignación jerárquica frente a punto, Datos numéricos y/o simbólicos, Determinística vs probabilística, Exclusivo vs superpuesto, Jerárquico vs plano, De arriba a abajo y de abajo a arriba, etc.

Existen diversos algoritmos de Clustering, entre los más conocidos están el Simple K-Means, El X-Means, Cobweb, El EM (Finite Mixture Models) dedicado al clustering probabilístico.

PREDICCIÓN

Técnica empleada en el data mining para proyectar los tipos de datos que se verán en el futuro o para predecir el resultado de un evento. Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo, los valores de las variables son generalmente continuos y las predicciones son usualmente sobre el futuro. Las variables pueden ser independientes, con atributos ya conocidos, o de respuesta, lo que queremos saber.

Las aplicaciones que se pueden encontrar en esta técnica son por ejemplo, revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro, predecir el precio de venta de una propiedad, predecir si va a llover en función de la humedad actual o por qué no, predecir la puntuación de cualquier equipo durante un partido de fútbol.

Las técnicas de predicción están basadas en modelos matemáticos y en ajustar una curva a través de los datos, esto se refiere a encontrar una relación entre los predictores y los pronosticados.

Todo basado en ajustar una curva a través de los datos, es decir, encontrar una relación entre los predictores y los pronosticados.

Las más comunes son: Modelos estadísticos simples como regresión, estadísticas no lineales como series de potencias, redes neuronales, etc.

Las redes neuronales utilizan los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión. Este proceso se conoce como entrenamiento de la red neuronal. Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.

PATRONES SECUENCIALES

La minería de datos secuenciales es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Son eventos que se enlazan con el paso del tiempo. El orden de acontecimientos es considerado. Por otra parte las reglas de asociación secuencial expresan patrones secuenciales, esto quiere decir que se dan en instantes distintos en el tiempo.

Entre las características de los patrones secuenciales están la importancia del cuerpo, su objetivo es encontrar patrones secuenciales, el tamaño de una secuencia es su cantidad de elementos, la longitud de la secuencia es la cantidad de ítems, el soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S, las secuencias frecuentes son las subsecuencias de una secuencia que tiene soporte mínimo.

Hablando de las ventajas tenemos que es flexible pues su comportamiento puede ajustarse gracias a su amplio conjunto de parámetros. Es eficiente ya que en cálculos muy sencillos, basta con recorrer una vez el conjunto de datos.

Por su contraparte en las desventajas se presentan la utilización que son los valores adecuados para los parámetros son difíciles de establecer a priori, por lo que se suele emplear un proceso de prueba y error. Sesgado por los primeros patrones, son los resultados obtenidos dependen del orden de presentación de los patrones.

En el proceso de los patrones secuenciales $|s|$es el número de elementos en una secuencia, una k-secuencia es una secuencia con k eventos. Una subsecuencia es una secuencia que está dentro de otra, pero que cumple ciertas normas. El ítem del evento i de la subsecuencia, está dentro del evento i de la secuencia.

VISUALIZACIÓN DE DATOS

La visualización de datos representa los datos en un formato ilustrado. Esto nos proporciona una manera accesible de comprender y entender los datos. Permite entenderlo de manera visual. Los tipos de visualización de datos que se presentan y mas frecuentes son los Gráficos: este tipo es el más común y conocido, se puede aplicar en hojas de cálculo como diagramas de árbol. Mapas: visualización de datos en mapas para poder visualizar sucesos en tiempo real como en los supermercados, cajeros automáticos, entre otros. Infografías: conjunto de imágenes, gráficos, texto simple que resume un tema para que se pueda entender fácilmente. Para procesar la información más compleja de una manera más fácil y entendible, Cuadros de mando: es una herramienta de gestión empresarial, es un conjunto de indicadores que aportan información para evaluar gestiones de compras, detectar amenazas y oportunidades.

Las diferentes aplicaciones que presenta son, comprender la información: mediante graficas de información, analizar y sacar conclusiones a partir de ese análisis. Identifica relaciones y patrones: se pueden vincular para reconocer parámetros con una correlación muy estrecha. Una gran cantidad de datos comienzan a tomar sentido y a su vez identificar tendencias emergentes: para descubrir tendencias en los negocios y mercados.

La visualización de datos es la mas importante medida de la era big data, es una herramienta cada vez mas importante para darle sentido a los millones de datos que se generan día con día, ayudando a presentar a estos mismos de una manera más fácil de comprender y leer, destacando tendencias y valores atípicos.

CLASIFICACIÓN

La clasificación es una técnica de las tareas predictivas, donde se predice el valor de un atributo basándose en los datos recolectados de otros atributos. Clasificación es el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene.

Empareja datos a grupos predefinidos, junta dependiendo del patrón que siguen los datos. Encuentra modelos que describen y distinguen clases o conceptos para futuras predicciones. La clasificación se considera como la técnica más sencilla y utilizada.

Los métodos utilizados son el análisis discriminante que se utiliza para encontrar una combinación lineal de rasgos que separan clases de objetos. Las reglas de clasificación esta busca términos no clasificados de forma periódica, para posteriormente si se encuentra una coincidencia se agrega a los datos de clasificación. Árboles de decisión, esté a través de una representación esquemática facilita la toma de decisiones. Solo puede tener un camino al cual seguir y las redes neuronales artificiales, que es un modelo de unidades conectadas para transmitir señales. Diferente a árbol de decisión tienes diversas respuestas.

Entre las características de estos métodos se encuentran la precisión en la predicción, la eficiencia, la robustez, la escalabilidad y la interpretabilidad.