UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

TÉCNICAS DE MÍNERIA DE DATOS

DESCRIPTIVAS Y PREDICTIVAS

LESLYE MARISOL HERNÁNDEZ BOLAÑOS.

1819111

GPO:03

PROF. MAYRA CRISTINA BERRONES REYES

MINERÍA DE DATOS

NUEVO LEÓN, MX, SEPTIEMBRE 2020

REGLAS DE ASOCIACIÓN

Esta técnica tiene muchas aplicaciones entre ella se encuentra el análisis de datos de la banca, otro ejemplo es que la gente que compra pan también comprará leche. Para obtener las reglas de asociación necesitamos de definiciones básicas como lo son:

- *Conjunto de elementos:* Una colección de uno o más artículos. Ejemplo: Leche, pan, mermelada.
- *Item set*: un conjunto de elementos que contiene k elementos.
- *Recuento de soporte:* frecuencia de ocurrencia de un ítem-set.
- Confianza (c): Mide que tan frecuencia del ítem en Y que aparecen en transacciones que contienen sigma elementos.

$$\frac{\sigma}{\#\ de\ transacciones}$$

Estrategias de generación de los elementos frecuentes. Un método para la generación de los elementos que aparecen con mayor frecuencia existe 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:

1. Identificar los ítems sets que ocurren con mayor frecuencia.
2. Convertir esos ítems sets frecuentes en reglas de asociación.

Otra estrategia para la generación de los elementos frecuentes es la “Class transformation”, esta consiste en cómo se escanean y analizan los datos, toda esta información almacenada contenida en el ítem está de manera vertical.

¿Cómo generar reglas? 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.

A continuación, haremos referencia al ejercicio de práctica con la estrategia Priori:

Lo primero es Identificar elementos frecuentes, posteriormente las ocurrencias, soporte y confianza. Para este ejercicio se consideró que el ítem set es frecuente si aparece un mínimo de 3 transacciones, es decir, su soporte debe ser igual o superior a 3/7 = .43. Además, hay que considerar que se inicia identificando todos los ítems individuales y recordar que las ocurrencias se toman como el número de veces que aparece el ítem en el elemento, para después obtener el soporte y analizar si este cumple está dentro del soporte que se estableció.

DETECCIÓN DE OUTLIERS / VALORES ATÍPICOS

Son observaciones de datos de los cuales sus valores son uy diferentes a las otras observaciones de este grupo de datos.

Muchos de los valores atípicos son ocasionados por:

- Valores de entrada de datos y procedimientos
- Acontecimientos extraordinarios
- Valores extremos y/o flotantes
- Causas no conocidas

Para calcularlos hay varios métodos, el más conocido es el método univariante, el cual, se centra en observar una sola característica del dato, mientras el método de multivariante toma en consideración diferentes características.

Técnicas

Existen diferentes técnicas para el cálculo de los atípicos, una de ellas es la Prueba de Grubbs, esta, plantea una hipótesis nula donde no hay valores atípicos, y una ha existe un valor atípicos, otra prueba conocida es la *Prueba de Dixon,* la cual, compara los valores extremos superior e inferiores y analiza la distancia entre ellos para posteriormente compararlos con los valores más centrados, por otro lado, tenemos la *Prueba de Tukey,* en donde se centra en el diagrama de caja, interpretando las líneas largas como presencia de valores atípicos, también existen técnicas computarizadas como los atípicos de *Mahalanobis*, o algunas otras que podemos obtener de forma manual calculando los mínimos cuadrado, como es el caso de la *Regresión Simple*.

Existen algunos programas con los cual podemos calcular los valores atípicos, como Excel o Minitab, cuando se encuentra un valor atípico se pueden sustituir o eliminarlo, pero antes hay que analizar si afecta o no a la muestra.

Aplicaciones

Tiene diferentes aplicaciones en el mundo real por ejemplo en la detección de fraudes financieros, como también en la tecnología informática y telecomunicaciones, así como en la nutrición y salud o en los negocios. Ejemplo; Un ejemplo claro de este tipo de técnica es analizar el desempleo a lo largo de los años de un grupo de personas, donde un valor atípico podría representar un periodo en el que, económicamente hablando, el país paso por un periodo bueno o malo, y eso afecto en la cantidad de desempleo por año.

REGRESIÓN LINEAL

Es un modelo matemático para determinar el grado de dependencia entre una o más variables, es decir, conocer si existe relación entre ellas.

Existen dos tipos de regresión:

Regresión lineal: Este hace referencia cuando una variable independiente ejerce influencia sobre otra variable dependiente.

Regresión múltiple: Se refiere a dos o más variables que ejercen influencia a una variable denominada como variable de respuesta.

En la minería de datos, la regresión lineal, está dentro de la categoría predictiva, ya que, analiza los datos y puede predecir qué pasará en un futuro, no quiere decir que es 100% precisa, solo nos muestra un pronóstico.

Análisis de regresión

Permite examinar la relación entre dos o más variables, las cuales, son las que tienen mayor impacto en un tema de interés. Algunos conceptos relevantes son los siguientes:

Variables dependientes: es la variable la cual es la más importante y la que se intenta entender o predecir el modelo.

Variable independiente: es el factor en el que se cree que puede impactar a la variable dependiente.

Ejemplo: Se cree que la cantidad de libra de vapor usadas en una planta por mes está relacionada con la temperatura ambiente promedio, para esto se hace una regresión lineal donde la temperatura es la variable dependiente y la cantidad de libra por vapor la variable independiente.

CLUSTERING

El clustering también es conocido como agrupamiento es una de las técnicas de minería de datos el proceso consiste en la división de los datos en grupos similares, se encargan de agrupar objetos, se mide que tan similares pueden ser entre ellas mismas y una vez que se realiza se colocan en grupos.

Cluster es una colección de datos que son similares entre sí, es decir, tienen una característica en común pero no son 100% iguales, con esto, podemos decir que el análisis de cluster se centra en encontrar las similitudes de los datos.

Aplicaciones

Estudios en terremotos: los epicentros observados deben agruparse a lo largo de las fallas continentales.

Aseguradoras: en la identificación de grupos de asegurados de seguros de automóviles con alto costo de promedio de reclamo.

En marketing ayudan a los profesionales de marketing a descubrir distintos grupos en sus bases de clientes.

Métodos de agrupación

- Asignación jerárquica frente a punto
- Datos numéricos y simbólicos
- Determinista vs probabilística
- Exclusivo vs super puesto jerárquico vs plano
- De arriba a bajo

El Simple k-means es un algoritmo que debe definir el número de clúster que se desean obtener así se convierte en un algoritmo voraz para participar, posteriormente se deberá determinar la cantidad de clusterings para después asumir de forma aleatoria los centros por cada cluster y así después quedan agrupados.

PATRONES SECUENCIALES

En la minería de datos secuencias, la extracción de patrones frecuentes están relacionados en tiempo además de que el orden es muy importante, además, se busca de la forma si sucede algo de la forma z en el instante t entonces sucederá el evento “y” en el tiempo t+n.

Características

- El orden es muy importante
- El objetivo de las secuencias es encontrar patrones secuenciales.
- La longitud de secuencia es la cantidad de ítems.
- El soporte es el % de secuencia que la contiene en un conjunto de secuencias
- Las secuencias frecuentes son las subsecuencia de una secuencia

Aplicaciones: Podemos diferenciar los patrones secuenciales con dos tipos diferentes con características básicas y de ahí proceder a observar las aplicaciones, estos tipos son los siguientes:

- *Agrupamiento de patrones secuencial:* Separa a grupos a los datos, de tal forma que, los grupos sean similares entre sí, pero diferentes de los otros grupos, por ejemplo, en la medicina, observar los patrones de si algún componente es causante de cáncer o de la misma manera podremos observar el comportamiento de las compras de los clientes en un supermercado.
- *Clasificación con datos secuenciales:* Se refiere a los datos contiguos que presentan un tipo de relación, por ejemplo, el reconocimiento de los correos que se mandan como spam debido a los caracteres que no son reconocidos con los anteriores.

Otros conceptos relevantes en esta técnica son las secuencias, las cuales, se representan con la letra S y representan el número de elementos en una secuencia, también este concepto esta acotado por los signos de <>, mientras los elementos se representan con los signos de {}, por otro lado, el término k-secuencial hace referencia a una secuencia con k eventos o también llamados ítems, también hay que considerar que, una subsecuencia es una secuencia dentro de otra pero cumpliendo ciertas normas que principalmente es el orden , y por último tenemos el término que hace referencia al ítem del elemento i de la tiene secuencia que esta adentro del evento i de una secuencia. Un ejemplo de esta técnica está en el considerar una trayectoria de un cliente que compra en una tienda, esto es la secuencia, los elementos estarían dados por los pasillos que recorre, mientras los ítems son los productos que toma de cada pasillo.

VISUALIZACIÓN

La visualización de datos nos sirve para representar gráficamente los elementos más importantes de la base de datos, que es, es la presentación de la información en formato ilustrado o gráfico. Es importante conocer cuál es el mejor gráfico para representar la información, los más comunes son los gráficos circulares, de líneas, barras aisladas, diagramas de dispersión, mapas u otro tipo de información como lo es Google maps, que nos ayuda a ver cosas que están pasando en tiempo real como en actividades de choque, también están las infografías, la cual, es una colección de imágenes gráficos que permite entender un tema fácilmente la información compleja, como también los cuadros de mando (dashboard), que son usados para las empresa ya que muestran los indicadores del negocio como ventas , proyecta a información en tiempo real para tomar decisiones adecuadas como colectivas.

Algunas técnicas más utilizadas para la creación de este tipo de gráficos están desde lo más sencillo como lo es Excel o algo más sofisticado como r.

Aplicaciones

- Comprender con sencillez gran cantidad de datos mediante el uso de representaciones graficas.
- Identificar tendencias, el uso de la visualización de datos para descubrir tendencia en los negocios y en el mercado puede dar una ventaja.
- Comunicar historia a otras personas ya que se puede resumir una información muy extensa.

Esta técnica es importante debido a que estamos en la era del big data por lo que la visualización es una herramienta cada vez más importante para darle sentido a los billones de datos que se generan cada día, la visualización ayuda a entender de forma más gráfica indicando la tendencia, los valores atípicos, etc.

CLASIFICACIÓN

La clasificación de tareas predictivas sirve para predecir el valor de un atributo en particular basándose en los datos recolectados de otros atributos.

Esta es una técnica de la minería de datos, ordena por clase tomando en cuenta las características de los elementos que lo contiene, de manera formal se puede definir como dada un base de datos D de tuplas (elementos o registros) con m clases, el problema de la clasificación es que trata de definir un mapeo f : D --> C donde cada tupla se le asigna una clase.

Datos de la clasificación empareja datos a grupos predefinidos además de que encuentra modelos que describen estos datos para futuras predicciones.

Métodos más utilizados

- Análisis discriminante: método utilizado para encontrar una combinación lineal como separar por colores.
- Reglas de clasificación: buscan términos no clasificados de forma periódica.
- Arboles de decisión: método analítico que a través de una representación esquemática que facilita la toma de decisiones.
- Redes neuronales artificiales: Es un modelo de unidades conectadas para trasmitir señales

Características

- Precisión en la predicción
- Eficiencia robustez y estabilidad en la interpretación

Un ejemplo de este tipo de técnica sería cuando los maestros clasifican a los estudiantes conforme a las calificaciones donde se le asigna una letra dependiendo su calificación.

PREDICCIÓN

Es una técnica que se utiliza para proyectar los tipos de datos que se utiliza para proyectar los tipos de datos que se verán en el futuro o predecir los resultados de un evento, por ejemplo, un partido de futbol que con frecuencia tiene tendencia histórica de ganar así que una predicción será el que gane en el próximo partido.

Existe cuestiones relativas a la relación temporal de las variables de entrada y las de salida, estas son lo que queremos llegar a saber

La predicción tiene relación con otras técnicas, pues, utiliza los datos históricos para observar el comportamiento de los datos.

Aplicaciones

- Predecir el precio de venta de una propiedad.
- Predecir si lloverá.
- Analizar el buró de crédito, para saber si te dan el crédito.

Técnicas

Se basan en modelos matemáticos, todo basado en ajustar una curva a través de los datos relación entre predictores y pronosticados, utiliza los tipos de métodos de regresión, es decir, la regresión lineal, la multivariable, la no lineal y la regresión no lineal multivariable.

Las redes neuronales también utilizan los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión.

Un ejemplo de este tipo de técnica es el pronóstico de ventas futuras de una empresa tomando en cuenta las ventas pasadas.