UANL

FCFM

Licenciatura en Actuaría

Minería de datos

Maestra Mayra Berrones

Ulises Solis Moises

1887850

02/10/20

-Reglas de asociación

Esta técnica funciona muy bien puesto que te sirve muy bien para encontrar patrones puesto que asocias los datos de maneras que puedes entender que tan frecuentes son y generas grupos óptimos.

Para obtener las reglas de asociación necesitamos de definiciones básicas como lo son:

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

Estrategias de generación de los elementos frecuentes

Uno de los métodos para la generación de los elementos que aparecen con mayor frecuencia que existen es el *Principio Priori*, este reduce el número de candidatos, si es frecuente entonces todos sus subconjuntos también serán frecuentes. Este es uno de los más comunes. Este algoritmo fue uno de los primeros en ser desarrollados y 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.

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.

-Detección de outliers

Esta técnica consiste principalmente en lo contrario a lo anterior, aquí revisamos datos que se encuentren fuera de los patrones, sería un dato que no forme parte del grupo de datos, o sea que no esté bien relacionado.

Valores atípicos

Son valores anormales comparados con el resto de datos en la base, no tienen su mismo comportamiento.

Los datos atípicos ocasionados por:

- *Errores de entrada y procedimiento*
- *Acontecimientos extraordinarios*
- *Valores extremos*

Existen distintos tipos de técnicas para detectarlos y se pueden dividir en dos categorías principales: Métodos univariables y métodos multivariables de detección.

TECNICAS PARA LA DETENCION de valores atípicos

- *Prueba de GRUBBS*
- *Prueba DIXION*
- *Prueba de TUKEY*
- *ANALISIS DE VALORES*
- *Regresión Simple*

Al detectar los outliers podemos eliminarlos o sustituir si son valores atípicos que no aportan nada, pero hay que realizarlo con cuidado ya que podemos sesgar la muestra y puede afectar al tamaño de la muestra, podemos afectar a la varianza de los datos lo que ocasionaría que nuestro análisis falle.

APLICACIONES DE OUTLIERS

- *Detección de fraudes financieros*
- *Tecnología informática y telecomunicaciones*
- *Nutrición y salud*

DISTINTOS SIGNIFICADOS

- *Error: Error a la carga de datos.*
- *Límites: Valores que se escapan de la media.*
- *Punto de interés: Casos anómalos que queremos detectar, ejemplo detectar cáncer.*

-REGRESION LINEAL RESUMEN

La primera vez que se usó algo parecido fue el método de mínimos cuadrados 1805. La primera vez que se usó formalmente y con la intención fue para revisar las estaturas y cómo influye la estatura de padres con los hijos.

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

- *Regresión lineal es cuando una variable influye a otra*

- *Regresión lineal múltiple es cuando unas variables influye a otra*

En la minería de datos se encuentra en la categoría de predictivo.

El análisis de regresión permite examinar la relación entre dos o más variables. Hay dos tipos de variables:

- *Variable(s) Dependiente(s): La variable que se intenta predecir.*
- *Variable(s) Independiente(s): Es el factor que influye en tu variable dependiente.*

Nos ayuda para poder predecir el futuro y mejoramiento de nuestras decisiones gracias a este análisis. Nos permite clasificar matemáticamente qué factores impactan más, cómo interactúan y cuánta seguridad nos brinda estos factores. Al mismo tiempo nos deja visualizar con muchos tipos de gráficos para entender la relación de estas variables. Este procedimiento nos va dando una serie de factores los cuales son los siguientes:

-La R representa el coeficiente de correlación y significa el nivel de asociación entre las variables.

-La R^2 representa el coeficiente de determinación, indica porcentualmente el cambio de la dependiente respecto a la independiente.

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 R^2 ajustada sea muy alta.

-CLUSTERING

Agrupamiento, que consiste en dividir todos los datos con características similares. Las diferentes técnicas usan algoritmos para poder dividir los datos.

Análisis de cluster

Dado un conjunto de puntos de datos tratar de entender su estructura. Encuentra similitudes entre los datos de acuerdo con las características encontradas en los datos. Es un aprendizaje no supervisado ya que no hay clases predefinidas.

APLICACIONES

- *Estudio de terremotos*
- Aseguradoras
- Planificación de una ciudad

Métodos de agrupación

- *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*

Algoritmos

- *Simple K-Means*
- *X-Means: Una mejora del K-Means*
- *Cobweb: Jerárquico*
- *EM: Finite Mixture Models*

-Predicción

Es una técnica que se suele usar para proyectar los tipos de datos, para predecir el resultado de un evento. Casi siempre el simple hecho de reconocer y comprender las tendencias históricas es suficiente para trazar una predicción un poco precisa de lo que podría ocurrir en el futuro.

Se tienen ciertas cuestiones relativas a la relación temporal de las variables de entrada o predictoras de la variable objetivo:

- *Los valores son generalmente continuos.*
- *Las predicciones son sobre el futuro.*
- *Las variables independientes corresponden a los atributos ya conocidos.*
- *Las variables de respuesta corresponden a lo que queremos saber.*

Aplicaciones

- *Banca: Saber cuánto se sacará de los cajeros*
- *Clima: Humedad cómo afecta*
- *Deportes: Predecir el resultado de cualquier equipo durante un partido de fútbol.*
- *Inmobiliaria: Predecir el precio de venta de una propiedad.*

Técnicas

Prácticamente las técnicas de predicción están basadas en modelos matemáticos y principalmente basados en ajustar una curva a través de los datos, esto se refiere a 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.

-Patrones secuenciales

Los patrones secuenciales se basan en el análisis de una secuencia y este con una base de datos ordenada por tiempo, o espacio, se busca encontrar un patrón que nos permita predecir el comportamiento con las características de tiempo o espacio.

Ventajas

- *Es flexible*
- *Es muy eficiente*

Desventajas

- *Difícil en ciertas ocasiones*
- *Sesgado por las primeras observaciones*

Tipos de datos

- *ADN y proteínas*
- *Recorrido de clientes en un supermercado*
- *Registros de usuarios en una página web*

Aplicaciones

- *Medicina: Predecir si un químico crea cáncer*
- *Márketing: Comportamiento de clientes*
- *Web: Reconocimiento de spam*

Posteriormente pasamos al análisis de la secuencia donde tenemos una base de datos, se escoge una secuencia, se selecciona un elemento y sus respectivos ítems para así evaluarlo.

El Método usado será el método GSP el cual funciona como el algoritmo a priori pero este genera las candidatas mediante la combinación de los más frecuentes de 1-secuencias y así tratándolos como si fuera la combinación de dos subsecuencias para generar una secuencia. Posteriormente se escoge los que superen el umbral y tengan la suficiente confianza y se escoge el mejor candidato.

-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.

Tipos

- Gráficos: más común, hojas de cálculo como diagramas de árbol, gráficos de dispersión etc.
- Mapas: visualización de datos en mapas para, para poder visualizar sucesos en tiempo real como google maps.
- Infografías: conjunto de imágenes, gráficos, texto simple que resume un tema para que se pueda entender fácilmente.
- Cuadros de mando: Cuadro de mando es una herramienta de gestión empresarial imprescindible e incluye indicadores.

Aplicaciones

- Comprender la información con rapidez
- Identifica relaciones y patrones
- Identificar tendencias emergentes

Importancia de la visualización de datos

Es muy importante visualizar los datos para poder manejar la información de manera eficiente y cada vez es una herramienta más importante para poder darle sentido a la gran cantidad de datos. Ayuda a contar historias de una manera más fácil de entender destacando tendencias y datos atípicos. Y también un punto muy fuerte debe ser darle la importancia necesaria a la tinta, el inkdata funciona que le des más tinta para los datos que realmente importan.

-Clasificación

Es una técnica de la minería de datos, también es el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene.

Datos de la clasificación

- 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.

Métodos utilizados

- Análisis discriminante: se utiliza para encontrar una combinación lineal de rasgos que separan clases de objetos.
- Reglas de clasificación: 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.
- Redes neuronales artificiales: modelo de unidades conectadas para transmitir señales. Diferente a árbol de decisión tienes diversas respuestas.

Características de los métodos

- Precisión en la predicción: Capacidad de predecir correctamente los datos
- Eficiencia: Realizar adecuadamente una función.
- Robustez: Habilidad de funcionar con ausencia de ciertos datos.
- Escalabilidad: Habilidad para trabajar con grandes cantidades de datos.
- Interpretabilidad: Facilidad de interpretación.