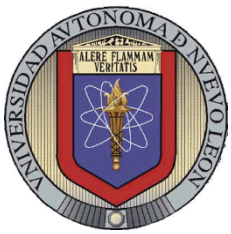


UNIVERSIDAD AUTONOMA DE NUEVO LEON

FACULTAD DE CIENCIAS FISICO MATEMATICAS

Minería de datos

TÉCNICAS DE MINERÍA

Laura Estefany Rodríguez de los Reyes
Matrícula: 1588292

MAESTRA: Mayra Cristina Berrones Reyes

REGLAS DE ASOCIACION

Las Reglas de asociación tiene la función de buscar 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 de información disponibles.

Algunas de sus aplicaciones son en el análisis de datos de la banca, el cross-marketing, el diiseño de catálogos, etc.

CONCEPTOS CLAVE

Soporte: es la fracción de transacciones que contiene un itemset.

Conjunto de elementos frecuente: es el conjunto de elementos cuyo soporte es mayor o igual que un umbral de mínimo.

Conjunto de elementos: colección de uno o más artículos,.

k- itemset, un conjunto de elementos que contiene k elementos.

Recuento de soporte: es la frecuencia de ocurrencia de un itemset.

Confianza (c): mide que tan frecuente los items en Y aparecen en transacciones que contienen X.

El objetivo de la minería de reglas de asociación es encontrar todas las reglas o patrones de una base de datos teniendo en cuenta lo siguiente:

- Umbral mínimo de soporte
- Umbral mínimo de confianza

Tipos de reglas de asociación:

- Enfoque en 2 pasos: se dice de dos pasos por que primero se generan elementos frecuentes y después se hacen reglas de asociación.
- Principio “a priori”: si un conjunto de elementos es frecuente, entonces todos sus subconjuntos son frecuentes.

CLASIFICACION

La clasificación es una técnica de la minería de datos, en la cual se ordena por clases tomando en cuenta las características de los elementos que contiene una base de datos especifica.

La técnica de clasificación tiene las sigueintes caracteristicas:

- Eficiencia
- Robustez
- Precisión en la precisión
- Interpretabilidad Datos de la clasificación
- Empareja todos los grupos predefinidos, junta dependiendo del patrón que siguen los datos
- Encuentra modelos que describen clases o conceptos para futuras predicciones
- La clasificación se considera como la técnica más sencilla.

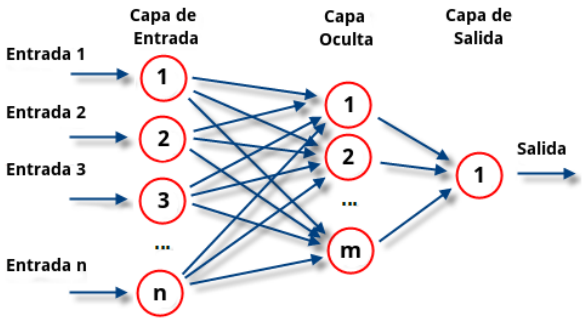
Algunos metodos para la clasificación:

Análisis discriminente: se utiliza para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos.

Reglas de clasificación: buscan términos no clasificados de forma periodica, si se encuentra una coincidencia se agrega a los datos de clasificación.

Arboles de decisión: permite determinar la decisión que se debe tomar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas . El árbol de decisión se construye partiendo el conjunto de datos en dos o más subconjuntos de observaciones, después estos subconjuntos se vuelven a particionar empleando el mismo algoritmo. La raíz del árbol es el conjunto de datos inicial, los subconjuntos y subsubconjuntos conforman las ramas del árbol. El conjunto en el que se realiza una partición se llama nodo y permite bifurcar en función de los atributos y sus valores. Las hojas del árbol proporcionan predicciones.

Redes neuronales artificiales: consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. La información de entrada atraviesa la red neuronal (donde se somete a diversas operaciones) produciendo unos valores de salida.



OUTLIERS

Outlier → Valor atípico

La detección de valores atípicos en el campo de la minería de datos y el descubrimiento de conocimiento a partir de datos es de gran interés en áreas que requieren sistemas de soporte a la toma de decisiones.

Los valores atípicos o también denominados anómalos tienen propiedades diferentes con respecto a la generalidad, ya que debido a la naturaleza de sus valores y por ende, a su comportamiento no son datos que mantienen un comportamiento similar a la mayoría. Esto ocasiona que los resultados se distorsionen.

Los datos atípicos son ocasionados por: errores de entrada de datos y procedimiento, acontecimientos extraordinarios, valores extremos y/o faltantes o por causas desconocidas.

Para calcular los valores atîicos hay diferentes métodos y se pueden clasificar en univariantes o multivariantes.

Algunas de estas técnicas para detectar los los valores atípicos son:

- Regresión simple
- Prueba de grubbs
- Prueba de Dixon
- Prueba de tuckey
- Análisis de valores y atípicos de mahalanobis

Hoy en día existen muchos programas que nos ayudan a detectar los valores atípicos, tales como: Rstudio, Minitab, Tableau, Excel y Google Analytics.

Entonces una vez que se detectaron los valores atípicos debemos eliminarlos o sustituirlos, aunque lo mejor sería quitarles valor a los datos atípicos con técnicas robustas.

La detección de outliers tiene muchas aplicaciones en la nutrición y area de la salud, la tecnologia informática y telecomunicaciones y en la detección de fraudes.

Los outliers o valores atípicos tienen distintos significados:

- Error: error a la carga de datos
- Punto de interés: casos anómalos que detectamos, por ejemplo, alguna enfermedad
- Limites: valores que son mas grandes o menores a la media

PREDICCIÓN

La predicción es una técnica que se utiliza para proyectar los tipos de datos que se verán en el futuro o predecir el resultado de un evento.

En muchos casos, el simple hecho de reconocer y comprender las tendencias históricas es suficiente para trazar una predicción algo precisa de lo que sucederá en el futuro.

La técnica de predicción tiene relación con otras técnicas ya que cualquiera de las técnicas utilizadas para la clasificación y la estimación puede ser adaptada para su uso en la predicción mediante el uso de ejemplos de entrenamiento donde el valor de la variable que se predijo que ya es conocido, junto con los datos históricos de esos ejemplos. Los datos históricos se utilizan para construir un modelo que explica el comportamiento observado en los datos. Cuando este modelo se aplica a nuevas entradas de datos, el resultado es una predicción del comportamiento futuro de los mismos.

Aplicaciones de la técnica de predicción:

- Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro.
- Predecir si va a llover mediante como ha llovido en años anteriores.
- Predecir eventos deportivos mediante resultados pasados.
- Predecir precios de propiedades basados en estudios del mercado.

La predicción hace uso de técnicas que se basan en modelos matemáticos como:

- Regresión simple
- Estadística no lineal
- Redes neuronales

Tipos de métodos de regresión:

- Regresión lineal
- Regresión lineal multivariante
- Regresión no lineal
- Regresión no lineal multivariante

REGRESIÓN LINEAL

Una regresión es un modelo matemático para determinar el grado de dependencia entre una o más variables, es decir conocer si existe relación entre ellas. Con esto podemos ajustar las variables a un modelo y así poder hacer una predicción de lo que puede pasar en un futuro. El objetivo del Análisis de regresión es determinar una función matemática sencilla que describa el comportamiento de una variable dados los valores de otra u otras variables.

Hay dos tipos de regresión:

• Regresión lineal: cuando una variable independiente ejerce influencia sobre una variable dependiente.

• Regresión lineal múltiple: Cuando dos o más variables independientes influyen sobre una variable dependiente.

El objetivo de la regresión es analizar los datos del conjunto y predecir lo que puede ocurrir en el futuro sin embargo no es muy preciso. Al mismo tiempo nos ayuda a visualizar con vario tipos de gráficos las relaciones de las variables utilizadas. 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 o que tanta relación lineal existe entre las variables.

La R^2 representa el coeficiente de determinación, indica porcentualmente el cambio que tiene la variable dependiente respecto a la o las variables independientes.

Se necesita saber si la regresión es significativa para tener idea si existe estas relaciones entre cada uno.

PATRONES SECUENCIALES

La técnica de patrones secuenciales es la extraccion de patrones que se presentan con frecuencia y se relacionan con el tiempo u otra secuencia. Son eventos que se enlazan con el paso del tiempo y el orden de acontecimiento se considera.

Por ejemplo en el caso de una tienda se tienen diferentes productos y existe el caso de que ciertos productos sean vendidos con regularidad en ciertas fechas, el sistema aprenderá de estos patrones. Dichos patrones serán importantes para poder utilizar la información de forma certera.

El objetivo es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos.

Ventajas:

▪ Es muy eficiente

▪ Flexible

Desventajas:

▪ Difícil en ciertas ocasiones

▪ Sesgado por las primeras observaciones

Las caracteristicas de esta técnica es que el orden de los datos importa, el tamaño de una secuencia es la cantidad de elementos y la longitud es la cantidad de items.

Algunos ejemplos de las aplicaciones que tiene son:

- Medicina (ADN y proteínas)
- Análisis de Mercado (comportamiento de compras)
- Web (spam en correo electrónico)

CLUSTERING

El clustering es una técnica en la cual se dividen los datos en grupos similares. Se utilizando algoritmos matemáticos se encargan de agrupar objetos. Se usa 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.

Cluster: colección de objetos de datos similares entre sí dentro de un grupo.

Análisis de cluster: cuando se da un cojunto de puntos se busca entender su estructura. Encuentra similitudes entre los datos de acuerdo con las caracteristicas que se encontraron.

Las aplicaciones que se pueden encontrar en el clustering son:

- Marketing: se busca encontrar distintos grupos en sus clientes.
- Planificación urbana: identica grupos de casas según su valor y ubicación geografica.
- Aseguradoras: buscar grupos de asegurados con un indice alto de reclamos.
- Estudio de terremotos: el epicentro del terremoto debe agruparse a lo largo de las fallas continentales.

Existen diferentes métodos de agrupación: asignación jerárquica frente a punto, determinística vs probabilística, datos numéricos y/o simbólicos, jerárquico vs plano y de arriba a abajo y abajo a arriba.

Algoritmos de clustering

- Simple K-Means: este tipo de algoritmo define el número de clusters que se desean obtener.

- X-Means: es una mejora del k-Means que es tener que seleccionar a priori el número de clustersque se desean obtener, a X-Means se le define un límite inferior k-min y un límite superior k-max y este algoritmo es capaz de obtener ese rango el número óptimo de clusters, dando de esta manera más flexibilidad.

- Cobweb: se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia.

- EM: pertenece a una familia de modelos que se conocen como Finite Mixture Models, los cuales se pueden utilizar para segmentar conjuntos de datos. Está clasificado como un método de particionado y recolocación

VISUALIZACIÓN DE DATOS

Esta técnica de la Minería de Datos representa gráficamente los elementos más importantes de una base de datos ya que se utilizan elementos visuales como cuadros, gráficos o mapas los cuales proporciona una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. La visualización de datos es la presentación de información en formato ilustrado o gráfico.

Tipos de visualización de datos:

▪ Gráficos: más común, hojas de cálculo como diagramas de árbol, gráficos de dispersión, etc.

▪ Mapas: visualización de datos en mapas para poder visualizar sucesos en tiempo real como Google maps.

▪ Cuadros de mando: cuando de mando es una herramienta de gestión empresarial imprescindible e incluye indicadores.

▪ Infografías: conjunto de imágenes, gráficos, texto siempre que resume un tema para que se pueda entender fácilmente.

Aplicaciones:

▪ Identifica relaciones y patrones

▪ Identificar tendencias emergentes

▪ Comprender la información con rapidez