Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Minería de datos

PRIMERA FASE

Resúmenes de Técnicas

de Minería de datos

Mayra Cristina Berrones Reyes

Alumna: Perla Millet Díaz Talamantes

Matrícula: 1809285

Grupo: 003

02 de octubre de 2020

REGLAS DE ASOCIACIÓN

En este tipo de técnica se describe una relación de asociación entre los elementos de un conjunto y tiene muchas aplicaciones, un ejemplo podría ser el análisis de datos de la banca, otro ejemplo es que la gente que compra harina para pastel también comprará decoraciones para el pastel.

CONCEPTOS BÁSICOS PARA ESTA TÉCNICA:

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

Ítem 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.

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

ESTRATEGIAS DE GENERACIÓN DE LOS ELEMENTOS FRECUENTES

el Principio Priori es un método para la generación de los elementos que aparecen con mayor frecuencia, 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.

Existe otra estrategia para la generación de los elementos frecuentes y 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 SE GENERAN LAS 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.

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

En este tipo de técnica se estudia el comportamiento de valores extremos que difieren del patrón general de una muestra.

LOS VALORES ATÍPICOS

Estos valores son diferentes a las observaciones del mismo grupo de datos.

Los datos atípicos son ocasionados generalmente por: Errores de entrada y procedimiento, acontecimientos extraordinarios o valores extremos.

Para la detección de estos valores, existen distintos tipos de técnicas y se pueden dividir en dos categorías principales: Métodos univariantes de detección y métodos multivariables

TECNICAS

- Prueba de GRUBBS
- Prueba DIXION
- Prueba de TUKEY
- Análisis de valores
- Regresión Simple

Cuando detectamos los outliers podemos eliminarlos o sustituir, pero hay que realizarlo con cuidado ya que podemos sesgar la muestra y puede afectar al tamaño de esta, también podemos afectar a la varianza.

APLICACIONES

Detección de fraudes financieros: cuenta que se abre y no tiene actividad en un gran tiempo y de repente recibe una fuerte cantidad de dinero

Tecnología informática y telecomunicaciones: detectar una falla del algoritmo que necesitamos procesar

Nutrición y salud: al tomar un grupo de personas con buena salud y puede ser un valor atípico alguien con presión alta.

Negocios: no puedes cambiar el giro del negocio con la información de dos outliers.

REGRESIÓN

Una regresión es un modelo matemático para determinar el grado de dependencia entre una o más variables, es decir, si existe relación entre ellas. En la minería de datos se encuentra en la categoría de predictivo.

Regresión lineal es cuando una variable influye a otra

Regresión lineal múltiple es cuando unas variables influyen a otra

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

- Variable dependiente: La variable que se intenta predecir
- Variable independiente: Es el factor que influye en tu variable dependiente

Este tipo de técnica nos ayuda para poder predecir el futuro y para el 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.

Aquí 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

Las técnicas de Clustering 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.

Un cluster es una colección de objetos de datos. Similares entre sí dentro del mismo grupo. Disimilar a los objetos en otros grupos.

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

Estudios de terremotos: los epicentros del terremoto observados deben agruparse a lo largo de fallas continentales.

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

Planificación de la ciudad: identificación de grupos de casas según su tipo de casa, valor, y ubicación geográfica.

Marketing: ayudar a los profesionales de marketing a descubrir distintos grupos en sus bases de clientes.

Uso del suelo: identificación de áreas de uso similar de la tierra en una base de datos de observación de la tierra.

MÉTODOS

- 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

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.

Un ejemplo claro que se tiene es respecto a un partido de fútbol, en el cual se tiene al equipo A contra el equipo B, y se observa que, de acuerdo a su tendencia histórica, el equipo A ha ganado el 80% de los partidos contra el equipo B, por lo que eso puede ser suficiente para predecir que el equipo A ganará el siguiente partido.

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 suelen ser sobre el futuro.
- Las variables independientes corresponden a los atributos ya conocidos.
- Las variables de respuesta corresponden a lo que queremos saber.

APLICACIONES

Banca: Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro.

Clima: Predecir si va a llover en función de la humedad actual.

Deportes: Predecir la puntuación 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

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. Se busca asociaciones de la forma “si sucede de la forma X en el instante de tiempo t entonces sucederá en el evento Y en el instante t+n”. El objetivo es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos.

CARACTERÍSTICAS

- El orden importa.
- Objetivo: 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.

APLICACIONES

- **Agrupamiento de patrones secuenciales**

  Medicina: Predecir si un compuesto químico causa cancer.
  Análisis de Mercado: Comportamiento de compras.

- **Clasificación con datos secuenciales**

  Web: Reconocimiento de spam de un correo electrónico.

VISUALIZACIÓN DE DATOS

Este tipo de técnica representa los datos en un formato ilustrado. Esto nos proporciona una manera accesible de comprender y entender los datos, permitiéndonos entender los datos de manera visual.

TIPOS

Gráficos: más común y conocido, en 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 en los supermercados, tránsito de vehículos, cajeros automáticos, etc. Un ejemplo es Google mapas.

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: Cuadro de mando es una herramienta de gestión empresarial imprescindible y es un conjunto de indicadores que aportan información para evaluar gestiones de compras, detectar amenazas y oportunidades.

APLICACIONES

Comprender la información con rapidez: mediante el uso de representaciones graficas de información para ver cantidades de datos en forma clara y cohesiva 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.

Identificar tendencias emergentes: el uso de visualización de datos para descubrir tendencias en los negocios y mercados

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.

MÉTODOS

Análisis discriminante: utilizado para encontrar una combinación lineal de rasgos que separan clases de objetos.

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

Árboles de decisión: a través de una representación esquemática facilita la toma de decisiones. Solo puede tener un camino al cual seguir.

Redes neuronales artificiales: es un modelo de unidades conectadas para transmitir señales. Diferente a árbol de decisión tienes diversas respuestas.

Características

- Precisión en la predicción: capacidad de predecir correctamente, grado de cercanía entre la precisión y el valor real.
- Eficiencia: realizar adecuadamente una función.
- Robustez: habilidad de funcionar con ausencia de ciertos valores.
- Escalabilidad: habilidad para trabajar con grandes cantidades de datos.
- Interpretabilidad: entendimiento que brinda.