Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

MINERÍA

Mayra Cristina Berrones Reyes

Resumen de las técnicas de minería de datos

Valeria Esthepania Urbina Gallegos 1799959

Lic. Actuaría Grupo: 003

San Nicolás de los Garza, Nuevo León 2 octubre 2020

Equipo 1: Reglas de asociación

Esta técnica tiene muchas aplicaciones.

Para comprender las reglas de asociación necesitamos de las siguientes definiciones:

Definiciones

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

Ítem set: 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.

Estrategias de generación de los elementos frecuentes

Un método para la generación de los elementos 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:

Etapas del Principio Priori

1. Identificar los ítems sets que ocurren con mayor frecuencia.
2. 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.

¿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.

Equipo 2. Detección de outliers

Estudia el comportamiento de valores extremos que difieren del patrón general de una muestra.

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

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

Técnicas de detección para los valores atípicos

- Prueba de GRUBBS
- Prueba DIXION
- Prueba de TUKEY
- Análisis 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.

Aplicaciones de outliers

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.

Distintos significados

Error: Error a la carga de datos

Límites: valores que se escapan de valores medios

Punto de interés: casos anómalos que queremos detectar como el ejemplo surgido en clase.

Equipo 3. Regresión lineal

La primera vez que se usó formalmente fue para revisar las estaturas y cómo influía la estatura de padres con la estatura de 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.

Tipo de regresiones

- Regresión lineal: una variable influye a otra
- Regresión lineal múltiple: diversas variables influyen a otra

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

Tipos de variables

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

Este método nos ayuda para poder predecir el mejoramiento de nuestras decisiones, 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.

Factores arrojados

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

Equipo 4: Clustering

Es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares.

Esta técnica es la más utilizada en algoritmos matemáticos se encargan de agrupar objetos.

Aplicaciones

Estudios de terremotos: los epicentros del terremoto observado deben agruparse a lo largo de fallas continentes

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

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.

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

Simple k-means

En este algoritmo se necesitan definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar.

Cobweb

Se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol donde las hojas representan los segmentos y el nodo rai engloba por completo el conjunto de datos.

Equipo 5: Predicción

Técnica que se utiliza para proyectar los tipos de datos, para predecir el resultado de un evento.

Relación con otras técnicas

La clasificación y la estimación pueden ser adaptadas para su uso en la predicción mediante el uso de ejemplos de entrenamiento.

Los datos históricos se utilizan para construir un modelo que explica el comportamiento observado en los datos.

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

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.

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

Equipo 6: Patrones Secuenciales

Conceptos

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.

Reglas de asociación secuencial: Expresan patrones secuenciales, esto quiere decir que se dan en instantes distintos en el tiempo.

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.

Ventajas:

- Flexibilidad: Su comportamiento puede ajustarse gracias a su amplio conjunto de parámetros.
- Eficiencia: Cálculos muy sencillos, basta con recorrer una vez el conjunto de datos.

Desventajas:

- Utilización: 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: Los resultados obtenidos dependen del orden de presentación de los patrones.

Aplicaciones

La clasificación con datos secuenciales: Donde datos contiguos presentan algún tipo de relación

Reconocimiento de caracteres escritos: Tiene como ejemplo la asociación de carácter a la identidad correspondiente entre un conjunto de símbolos que componen el alfabeto, ayuda a automatizar la lectura de direcciones postales, cheques bancarios, formularios.

Agrupamiento de patrones secuenciales: separa en grupos a los datos, encuentra agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos.

Análisis de transacciones del cliente: en este caso, las transacciones representan conjuntos de elementos que coexisten en el comportamiento de compra del cliente. En este caso, es conveniente. Determinar patrones frecuentes de comportamiento de compra, ya que pueden utilizarse para tomar decisiones sobre las existencias en los estantes o las recomendaciones.

Equipo 7: 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 de visualización de datos

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.

Aplicaciones

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.

Identificar tendencias emergentes: para descubrir tendencias en los negocios y mercados

Equipo 8. 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

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