UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO-MATÉMATICAS

Unidad de Aprendizaje:

Minería de Datos

Resumen de las Técnicas de Minería de Datos

Semestre Agosto 2020 – Enero 2021

Maestro(a):

Mayra Cristina Berrones Reyes

Alumno(a):

Andrea López Solís 1822031

Semestre: 7to.

Grupo: 003

Licenciatura en Actuaria

Monterrey, Nuevo León, 2 de octubre de 2020

TÉCNICAS DE MINERÍA DE DATOS – DESCRIPTIVAS

CLUSTERING

El clustering también llamado agrupamiento es un proceso que consiste en la división de los datos en un grupo de objetos similares. Este proceso se lleva a cabo mediante la información que brindan las variables que pertenecen a cada objeto, se mide la similitud entre los mismos, y una vez ocurrido eso se colocan en clases que son similares internamente y a la vez diferente entre los miembros de las diferentes clases. El clustering es un aprendizaje no supervisado ya que no hay clases predefinidas.

Aplicaciones del clustering:

- Estudios de terremotos: los epicentros del terremoto observados deben agruparse a lo largo de fallas continentales.
- Aseguradoras: identificación de grupos de aseguradoras de seguros de automóviles con 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 tierra.
- Marketing: ayuda a los profesionales de marketing a descubrir distintos grupos en sus bases de clientes.
- Planificación de la ciudad: identificación de tipo de casas según su tipo de casa, valor, ubicación geográfica.

Métodos de Agrupación que se utilizan en el clustering:

- Asignación jerárquica frente a punto.
- Datos numéricos y/o simbólicos.
- Determinística vs. Probabilística.
- Exclusivo vs. Supuesto.
- Jerárquico vs. Plano.
- De arriba a bajo y viceversa.

Algoritmos de Clustering

- **Simple K-Means**:se debe definir el número de clusters que se desean obtener.

- **X-Means**: Algoritmo mejorado del K-Means. Este algoritmo define un límite inferior K-min (núm. mínimo de clusters) y un límite superior K-Max (núm. máximo de clusters), este algoritmo es capaz de obtener en este rango el número óptimo de clusters, siendo de esta manera más flexible.
- **Cobweb**:Algoritmo jerárquico que utiliza aprendizaje incremental (realiza las agrupaciones instancia a instancia). Toma en consideración dos parámetros muy importantes: **Acuity** (es un parámetro que está basado en la estimación de la media y la desviación estándar del valor de un atributo para un nodo en particular) y **Cut-off** (es un parámetro usado para evitar el crecimiento descontrolado de la cantidad de segmentos, indica el grado de mejoría que se produce en la utilidad de categoría).
- **EM:** Se utiliza para segmentar conjuntos de datos, es un método de particionado y recolocación, o sea, Clustering Probabilístico. El algoritmo EM, procede en dos pasos que se repiten de forma iterativa: **Expectation y Maximization**.

REGLAS DE ASOCIACIÓN

Las reglas de asociación son búsquedas de 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. En términos más simples las reglas de asociación en sí describen una relación de asociación entre los elementos de un conjunto de datos relevantes.

Algunas aplicaciones de esta técnica serian:

- Análisis de datos de la banca.
- Cross-marketing (poner la crema batida junto a las fresas).
- Diseño de catálogos.

Dos conceptos de suma importancia en la técnica de reglas de asociación son:

- Soporte: El porcentaje de las transacciones que contienen todos los ítems de X e Y (donde X es denominado el antecedente de la regla e Y su consecuente).

  El Soporte se puede apreciar como una probabilidad (X -> Y) donde:

$$s = \frac{\sigma(\{X, Y\})}{\#\text{Núm. transiciones}}$$

- Confianza: Mide que tan frecuentes ítems en Y aparecen en transacciones que contienen X.

  La confianza se puede apreciar como si fuera una probabilidad condicional, donde tenemos:

$$c = \frac{\sigma(\{X, Y\})}{\sigma(\{PanY\})}$$

Dado un conjunto de transacciones T, el objetivo de la minería de reglas de asociación es encontrar todas las reglas teniendo: **Umbral mínimo de soporte** y **Umbral mínimo de confianza**.

En las reglas de asociación se toma en cuenta el Principio de "A priori", el cual establece que, si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes.

El algoritmo "A priori" fue uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación y sigue siendo uno de los más empleados, consta de dos etapas:

- Identificar todos los itemsets que ocurren con una frecuencia por encima de un determinado límite (itemsets frecuentes).

    1. Se calcula el soporte de cada ítem individual, y se establecen los 1eros-itemsets frecuentes.
    2. Los itemsets generados anteriormente se vuelven a utilizar para generar los nuevos itemsets candidatos.
    3. Se calcula el soporte de cada itemset candidato y se determinan los itemset frecuentes que serán aquellos cuyo soporte sea mayor o igual al mínimo establecido.
    4. El proceso sigue hasta que la condición del soporte no se cumpla y ya no puedan ser encontrados nuevos itemsets frecuentes.

- Convertir esos itemsets frecuentes en reglas de asociación.

    Una vez encontrados todos los itemsets frecuentes se procede a verificar la confianza para solo tomar en consideración aquellos cuya confianza sea mayor o igual a la confianza mínima supuesta.

DETECCIÓN DE OUTLIERS

La detección de outliers estudia el comportamiento de valores externos que difieren del patrón general de una muestra. El objetivo de esta técnica es encontrar patrones que ven un resumen de las relaciones ocultas que existen entre los datos.

Para entender la detección de outliers hay que considerar lo que es son los valores atípicos.

Un valor atípico son observaciones cuyos valores son diferentes a las otras observaciones del mismo conjunto de datos. Estos datos pueden llegar a distorsionar los resultados de los analisis por ello es importante identificarlos y tratarlos adecuadamente, ya sea que se eliminen o sustituyan.

Estos datos se pueden calcular mediante distintos tipos de técnicas que se dividen en 2 categorías: **Métodos univariantes y métodos multivariantes**.

Algunas técnicas para detectar los valores atípicos son:

- Prueba de Grubbs.
- Prueba de Dixon.
- Prueba de Tukey.
- Analisis de valores atípicos de Mahalonobis.
- Regresión simple.

Aplicaciones de la minería de datos en outliers:

- Detección de fraudes financieros.
- Tecnología informatica y telecomunicaciones.
- Nutrición y salud.
- Negocios.

Un Outliers puede ser:

- Error.
- Límites: un dato que se escapa de un “grupo medio”, queremos mantener el dato modificado, para que no perjudique el aprendizaje del módulo de ML.
- Punto de interés: casos “anómalos” los que queremos detectar y que sean nuestro objetivo.

VISUALIZACIÓN

La visualización de datos es la presentacion de información en un formato ilustrado o gráfico. Es el proceso de búsqueda, interpretación y comparación de datos que permite un conocimiento profundo de los mismos datos de tal forma que estos sean útiles para el usuario en un futuro.

Tipos de visualización de datos:

- Gráficos: Para representar datos de manera sencilla, como gráficos circulares, líneas, columnas, barras, burbujas, diagramas de dispersión y mapas de tipo árbol.
- Mapas.
- Infografías: ayudan a procesar más fácil la información compleja.
- Cuadros de Mando (Dashboards): es una herramienta que permite saber en todo momento el estado de los indicadores del negocio.

Técnicas para la visualización de datos

La mayoría de los analistas utilizan software avanzados para analizar y visualizar datos. Las herramientas de software van desde hojas de cálculo sencillas con Excel o Google Sheets a software de analítica más sofisticado como R.

Aplicaciones de la técnica de visualización:

- Comprender la información con rapidez.
- Identificar relaciones y patrones.
- Identificar tendencias emergentes.
- Comunicar la historia a otras personas.

La visualización de datos es importante a medida que la “era del big data” entra en pleno apogeo, la visualización le da sentido a los billones de filas de datos que se generan cada dia. La visualización de datos ayuda a contar historias seleccionando datos en una forma más fácil de entender, destacando las tendencias y los valores atípicos. Una buena visualización cuenta una historia eliminando el ruido de los datos y resaltando la información útil.

TÉCNICAS DE MINERÍA – PREDICTIVAS

REGRESIÓN

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

Existen dos **tipos** de regresión:

1. Regresión Lineal: cuando una variable independiente ejerce influencia sobre otra variable dependiente.
2. Regresión Lineal Múltiple: cuando dos o más variables independientes influyen sobre una variable dependiente.

En la minería de datos la Regresión es una técnica que se encuentra dentro de la categoría de técnicas de minería predictivas.

Análisis de Regresión

El analisis de regresión permite examinar la relación entre dos o más variables e identifica cuales son las variables que tienen mayor impacto en un tema de interés. La regresión toma en consideración dos variables que son:

- Variable dependiente: es el factor el cual se está tratando de entender.
- Variable independiente: es el factor que tú crees que puede impactar a la variable dependiente.

El analisis de regresión nos permite explicar un fenómeno y predecir situaciones futuras, por lo que es de ayuda para toar decisiones y asi obtener mejores resultados.

CLASIFICACIÓN

La clasificación es una técnica predictiva de la minería de datos. Es el ordenamiento o disposición por clases tomando en cuenta las caracteristicas de los elementos que contiene.

Algunos **métodos de clasificación** son:

- Analisis discriminante: método usado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos. Es uno de los métodos más sencillos y fáciles de comprender. Un ejemplo de este sería la separación de colores o calificaciones.
- Reglas de clasificación: buscan términos no clasificados de forma periódica. Un ejemplo de este método son los algoritmos usados en YouTube donde aparece contenido relacionado dado las búsquedas previas.
- Árboles de decisión: método analítico que facilita la forma de decisiones. Los árboles de decisiones solo tienen una respuesta o un solo camino a seguir. Un ejemplo aplicado de este método lo podemos ver en los algoritmos de programación donde se usa el condicional IF-ELSE.
- Redes neuronales artificiales: es un modelo de unidades conectadas para transmitir señales. Este método es parecido a los árboles de decisión solo que a diferencia este tiene más respuestas o caminos de decisiones a tomar, es decir, que una sola decisión puede tener múltiples caminos.

Caracteristicas de los métodos de clasificación:

- Precisión en la predicción.
- Eficiencia.
- Robustez.
- Escalabilidad.
- Interpretabilidad.

PATRONES SECUENCIALES

Los patrones secuenciales se explican mediante dos conceptos:

- Minería de Datos secuenciales: extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Son eventos que se enlazan con el paso del tiempo, aquí es importante considerar el orden.
- Reglas de asociación secuencial: expresan patrones secuenciales, quiere decir que son sucesos que se dan en instantes distintos en el tiempo.

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

Caracteristicas de los patrones secuenciales:

- Es importante tomar en consideración el orden.
- 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 secuencia que la contiene en un conjunto de secuencias S.
- Las secuencias frecuentes son las subsecuencias de una secuencia que tiene soporte mínimo.

Aplicaciones de los patrones secuenciales:

- Medicina: Predecir si un compuesto químico causa cáncer.
- Análisis de Mercado: Comportamiento de compras.
- Web: Reconocimiento de spam de un correo electrónico.

Para **realizar patrones secuenciales** se toman en cuenta las **fases del método GSP** (Generalized Sequential Pattern) que consta de dos fases:

1. Recorre las bases de datos para obtener todas las secuencias frecuentes de 1 elemento.
2. Generar k-secuencias candidatas a partir de las (k-1)-secuencias frecuentes.
   Podar k-secuencias candidatas que contengan algunas (k-1)-secuencia no frecuente.
   Conteo: obtener el soporte de las candidatas.

Eliminar las k-secuencias candidatas cuyo soporte real este por debajo del umbral de soporte mínimo frecuente.

- PREDICCIÓN

Esta es una técnica que se utiliza para proyectar los tipos de datos que se verán en un futuro o predecir el resultado de un evento. Tan solo con reconocer y comprender las tendencias históricas es suficiente para trazar una predicción algo precisa de lo que sucederá en el futuro.

Caracteristicas de la técnica de predicción:

- Los valores son generalmente continuos.
- Las predicciones son a menudo sobre el futuro.

Esta técnica se puede relacionar con otras técnicas, ya que cualquiera de las técnicas utilizadas para la clasificación y la estimación puede ser adaptada para su uso en las predicciones. Los datos históricos se pueden utilizar también 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 de comportamiento futuro de los mismos datos.

Aplicación de la técnica de Predicción:

- Revisar los historiales crediticios de los consumidores para predecir si serán un riesgo crediticio en el futuro.
- Predecir el precio de venta de una propiedad.
- Predecir si va a llover en función de la humedad actual.
- Predecir la puntuación de cualquier equipo durante un partido de fútbol.

La mayoría de las técnicas de predicción se basan en modelos matemáticos, todo basado en ajustar una curva a través de los datos (es decir encontrar una relación entre los predictores y los pronosticados):

- Modelos estadísticos simples como regresión.
- Estadísticas no lineales como series de potencias.
- Redes neuronales, RBF, etc.

Tipos de métodos de regresión: Regresión Lineal, Regresión lineal Multivariante, Regresión No Lineal y Regresión No Lineal Multivariante**.**

Redes neuronales: Usa datos para modificar las conexiones ponderadas entre sus funciones hasta que es capaz de predecir los datos con precisión.