UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

ARMANDO JAVIER DELGADO CANTU

1887833

Clase: MINERIA DE DATOS

Maestra: Mayra Berrones

“RESUMENES DE LAS DIFERENTES TECNICAS DE MINERIA DE DATOS”

Octubre del 2020

Clustering

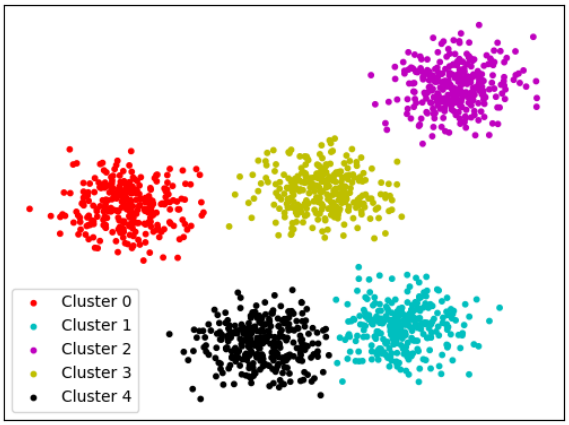
El término clustering consiste en dividir los datos en grupos de objetos similares. Las técnicas de clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información de las variables que pertenecen a cada objeto se mide la similitud o parecido entre los mismos, 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.

Los objetos se agrupan en distintos ‘’clúster’’. Un clúster es una colección de objetos de datos que son similares entre sí dentro del mismo grupo. Después se analiza cada uno para entender la estructura y encontrar similitudes entre los datos recolectados

Esta técnica tiene distintas aplicaciones como, por ejemplo:

- Estudios de terremotos y sus epicentros
- Planeación geográfica
- Marketing para entender a los distintos grupos de clientes
- Planificación de la ciudad y ubicaciones geográficas
- En las aseguradoras para los distintos tipos de clientes y seguros

Algunos algoritmos famosos de clustering son los siguientes: Simple K-Means, X-Means, Cobweb.



Reglas de Asociación

Proceso que se utiliza mucho en la inteligencia artificial que busca un descubrimiento de tendencias o patrones en bases de datos muy amplias. En general, lo que se trata de hacer es describir una regla de asociación entre los elementos de un conjunto de datos relevantes. Es la búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de elementos u objetos en bases de datos.

Sus aplicaciones pueden servirnos de muchas maneras, desde el análisis de datos de la banca, hasta el cross-marketing y el diseño de catálogos. Un ejemplo muy fácil de entender es por ejemplo es por ejemplo cuando en una tienda ponen estratégicamente las galletas a lado del café o así, para que la gente compre ambas ya que las galletas y el café se asocian porque la gente gusta de comerlos juntos.

Para las reglas de asociación, es importante destacar que la confianza con la que se mide la frecuencia de un ítem ‘Y’ aparece en transacciones del ítem ‘X’, 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. El enfoque de fuerza bruta, consiste en que teniendo listas todas las reglas de asociación, comprobando el soporte y la confianza, se eliminen las reglas que fallan según los umbrales mínimos de confianza y de soporte.

Las estrategias de generación de los elementos frecuentes 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. El primero es Identificar los ítems sets que ocurren con mayor frecuencia y el segundo convertir esos ítems sets frecuentes en reglas de asociación.

Clasificación

Es una técnica de la minería de datos que se utiliza para el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene cierta base de datos. La clasificación se encuentra dentro de las técnicas predictivas ya que trata de acertar a lo que va a pasar en un futuro de un atributo en particular basándose en los datos recolectados de otros atributos. La clasificación empareja datos a grupos predefinidos, junta dependiendo del patrón que siguen los datos. Se encuentran modelos que describen y distinguen clases para llevar a cabo predicciones a futuro. La clasificación se considera como la técnica de minería de datos más sencilla y utilizada. Esto es por su simplicidad para realizarse y sus resultados eficientes.

Sus principales características son las siguientes

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

Un ejemplo puede ser en los equipos de futbol en un torneo, se pueden agrupar dependiendo de sus fortalezas o debilidades.

Los métodos más utilizados en la clasificación son:

- El análisis discriminante: se usa para encontrar una combinación lineal de rasgos que separan clases de objetos
- Reglas de clasificación: busca periódicamente términos no clasificados, para luego si se encuentra una coincidencia se agrega a los datos clasificados
- Árboles de decisión: Por medio de una representación esquemática simplifica la toma de decisiones. Solo hay un camino para seguir
- Redes neuronales artificiales: Es un modelo de unidades conectadas para transmitir señales.

Outliers

La detección de outliers o valores atípicos, pertenece a la categoría de técnicas descriptivas y su función es estudiar el comportamiento de valores externos que difieren del patrón general de una muestra. Los Valores atípicos son valores muy diferentes a las observaciones del mismo grupo de datos. Los datos atípicos son principalmente ocasionados por errores de entrada y procedimiento, acontecimientos extraordinarios, valores extremos o por causas no conocidas.

Los datos atípicos distorsionan los resultados de los análisis por lo que hay que identificarlos y manejarlos de manera apropiada, sino nos puede dar como consecuencia un mal análisis estadístico o una mala predicción.

Existen varios tipos de técnicas para detectar outliers (datos atípicos) y se pueden dividir en dos categorías principales las cuales son:

- Métodos univariantes de detección
- Métodos multivariables

Las principales técnicas para la detección de valores atípicos son la prueba de GRUBBS, prueba DIXION, prueba de TUKEY, análisis de Valores y regresión simple.

Por ejemplo, vamos a tomar un conjunto de datos que representa las temperaturas de 12 objetos diferentes en un cuarto. Si 11 de los objetos tienen temperaturas cercanas a 21 grados Celsius, pero el duodécimo objeto, un horno, tiene una temperatura de 150 grados Celsius, una observación rápida te indicará que probablemente el horno sea un valor atípico. Esto es porque ese dato extremo puede alterar el promedio de temperaturas.

Regresión

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.

Existen dos tipos de regresión:

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

En Minería de Datos la Regresión se encuentra dentro de la categoría predictivo. Esta categoría tiene como objetivo analizar los datos de un conjunto y en base a eso, predecir lo que puede ocurrir con ese conjunto de datos en un futuro.

El análisis de regresión permite examinar la relación entre dos o más variables e identificar cuáles son las que tienen mayor impacto en un tema de interés. Existen 2 tipos de variables a la hora de hacer regresión:

- Variable(s) dependiente(s): Es el factor más importante, el cual se está tratando de entender o predecir.
- Variable(s) independiente(s): Es el factor que tú crees que puede impactar en tu variable dependiente.

El **análisis de regresión** nos permite explicar un fenómeno y predecir cosas acerca del futuro, por lo que nos será de ayuda para tomar decisiones y obtener los mejores resultados.

Visualización de Datos

La visualización de datos es una manera de representar los datos en un formato ilustrado para facilitar su comprensión visual. Esto nos proporciona una manera accesible de entender los datos de una manera más dinámica.. Los tipos de visualización de datos que se presentan y más frecuentes son los siguientes:

- Gráficos: este tipo es el más común y conocido, se puede aplicar en hojas de cálculo como diagramas de árbol (a estas alturas todos conocemos los gráficos)
- Mapas: visualización de datos en mapas para poder ver sucesos en tiempo real como en los supermercados, cajeros automáticos, centros comerciales, entre otros.
- Infografías: Son un conjunto de imágenes, gráficos, o hasta texto simple que resume un tema para que se pueda entender fácilmente. Procesa la información para que se entienda de una manera simplificada.
- Cuadros de mando: Es una herramienta de gestión empresarial, son un conjunto de indicadores que aportan información para evaluar gestiones de compras, detectar amenazas y oportunidades.

Las diferentes aplicaciones que tiene son las siguientes:

- Comprender la información con rapidez: Se puede analizar una gran cantidad de información importante de manera sencilla.
- Identificar relaciones y patrones: A veces ver la información de muchos datos en una gráfica puedes observar la relación que existe entre ellos o algún patrón que se repita en la gráfica.
- Identificar tendencias emergentes: A través de las gráficas se pueden ver las tendencias que tiene algún activo y así sacarle ventaja al mercado.

Patrones Secuenciales

Un concepto que tenemos que saber antes es la minería de datos secuenciales que es la extracción de patrones frecuentes relacionados con el tiempo o algún otro tipo de secuencia. Son eventos que se enlazan con el paso del tiempo por su ocurrencia.

Otro termino importante son las reglas de asociación secuencial que son las que expresan patrones secuenciales, esto nos dice que se dan en instantes distintos en el tiempo. Entre las características de los patrones secuenciales están la importancia del cuerpo, su objetivo es 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, es flexible pues su comportamiento puede ajustarse gracias a su amplio conjunto de parámetros. Es eficiente por su simplicidad para recorrer una vez el conjunto de datos.
- Desventajas: Utilización que son los valores adecuados para los parámetros son difíciles de establecer a priori, por tanto, se suele emplear un proceso de prueba y error. Es sesgado por los primeros patrones, son los resultados obtenidos dependen del orden de presentación de los patrones.

Se puede utilizar en distintas áreas por ejemplo en comportamiento de compras de algún cliente o consumidor en un sitio web o mercado.