UNIVERSIDAD AUTÓNOMA
DE NUEVO LEÓN

FACULTAD DE CIENCIAS
FISICO MATEMÁTICAS

Licenciatura en Actuaría

Minería de Datos

Grupo 003

Resúmenes técnicas de minería

Docente Mayra Cristina Berrones Reyes

Alumna

Estibalyz Villarreal Martínez

Matrícula 1809399

2 de octubre 2020

Técnicas de minería de datos

Reglas de Asociación

Es la búsqueda de patrones frecuentes, casualidades, correlaciones y cosas similares en una base de datos, todo esto se asemeja a un tema de la materia de portafolios de inversión ya que de la línea de cómo van las acciones tenía varias frecuencias y “picos” ya sea a la alza o a la baja de los cuales nosotros podíamos ver una tendencia para sacar conclusiones de cómo se iba a comportar el mercado al igual si estamos analizando un año, el cómo se comportaría el siguiente año.

Las reglas de asociación están en todos lados, es muy interesante el cross-marketing ya que desde tiendas muy pequeñas hasta imperios lo usan, el tener juntos los productos que puede que se compre en la misma ida a la tienda, un ejemplo es el pan y la mermelada, la mayoría de las veces están juntos. El diseño de catálogos sigue una regla de asociación, en una parte te pondrá electrodomésticos, en otra parte te pondrá cosas para la jardinería, etc. El objetivo de las reglas de asociación es que habiendo un recorrido, un conjunto de transacciones T, es tener todas las reglas teniendo el umbral mínimo de soporte y el umbral mínimo de confianza.

Si hablamos de un enfoque en 2 pasos cada regla va a ser una partición binaria de un conjunto de elementos frecuentes. Al buscar las reglas de asociación lo más común que se encuentra es el principio “apriori”, este dice que si un conjunto de elementos es frecuente, entonces sus subconjuntos deben ser frecuentes, este se basa en comprimir una gran base de datos en una estructura compacta de árbol de patrones frecuentes, evitando así costosos análisis de bases de datos. El soporte de un conjunto de elementos nunca excede el soporte de sus subconjuntos, esto se conoce como la propiedad anti-monótona de soporte.

Outliters

En español “outliters” se puede traducir como “parte aislada”. La detección de outliters es el estudio del comportamiento de valores extremos que difieren del patrón general de una muestra, aquí toman papel los valores atípicos, estos son los que toman valores muy diferentes a las otras observaciones del mismo grupo, hay varias razones por las que pueden ser ocasionados, por errores de entrada de datos, acontecimientos extraños, valores extremos o faltantes y causas no conocidas. Los datos atípicos distorsionan los resultados de los análisis por lo que tienen que ser tratados. Esto pasa en la línea de tendencia de las acciones, puede que porque un día amanece el dólar con una dispersión considerada, si muchos venden sus acciones o si muchos compran, aprovechando la situación, ese día se verá un dato atípico por un acontecimiento extraño.

Existen técnicas para detectar los valores atípicos, se pueden dividir en grandes ramas métodos univariantes de detección de outliters y métodos multivariantes de detección de outliters, las técnicas son Prueba de Grubbs, Prueba de Dixon, Prueba de Tukey (diagrama de caja), Análisis de Valores (Atípicos de Mahalanobis), Regresión Simple (regresión por mínimos cuadrados). Hay aplicaciones para identificar los outliters como son R, Google Analytics, Minitab, Tableau y como más sencillo y básico está Excel.

Ya que son detectados los outliters se pueden sustituir o eliminar si se trata de un error en la captura; si no es un error, la mejor opción es quitarle peso a esas observaciones ya que si se elimina introduce un sesgo, disminuye el tamaño muestral y puede afectar la distribución y varianza.

Patrones Secuenciales

Para entender los patrones secuenciales se debe saber a lo que se dedica la minería de datos secuenciales, es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia.

Para encontrar patrones secuenciales hay que tener claras las características como que el orden importa, el tamaño de una secuencia es su número de elementos sin omitir alguno, la longitud de la secuencia es el número de ítems, el soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S y que las secuencias frecuentes son las subsecuencias de una secuencia que tiene soporte mínimo. Una subsecuencia es una secuencia que está dentro de otra, pero cumpliendo ciertas normas, una de ellas es que el ítem del evento i de la subsecuencia, tiene que estar dentro del evento i de la secuencia, también una regla es que debe pasar por un método de eliminación, cuando el soporte real esté por debajo del umbral de soporte mínimo de frecuencia.

Las ventajas de los patrones secuenciales es que son fáciles de usar y aplicar, como desventaja es que ocurre un sesgo con los primeros patrones. Las aplicaciones en la minería de datos se dividen en dos, que es el agrupamiento de patrones secuenciales y la clasificación con datos secuenciales. Por ejemplo en el agrupamiento, en la medicina que se puede predecir si un tipo de compuesto químico causa cáncer. Otro ejemplo de agrupamiento sería el análisis en el mercado, para ser más específicos, las compras, aquí todo está relacionado, desde que lo que esta acomodado en fila lo primero tiene fecha de caducidad más pronta que los de atrás, también están acomodados los productos por lo que casi siempre se compra en conjunto. En la clasificación de los datos secuenciales podríamos decir que el mismo centro de mensajes reconoce cuando un correo es spam y lo manda a otra carpeta que no es la principal.

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, como es el caso de las acciones, las cuales las podemos encontrar gratis en algún modelo de finanzas como Yahoo! Finanzas su línea de tendencia, y de ahí sacar conclusiones de cómo se comportara en semanas, meses o años.

Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo. Hay variables independientes que son los atributos ya conocidos y variables respuesta que es lo que queremos saber.

Las aplicaciones de la predicción son múltiples, predecir el precio de venta de una casa, terreno, carro; predecir si va a llover en función a la humedad actual; predecir la puntuación de cualquier equipo durante un partido de fútbol; revisa el historial del cliente para ver si al sacar algo a crédito lo pueda pagar.

Las técnicas se basan en una ajustar una curva a través de los datos, encontrando una relación con los predictores y pronosticados. La mayoría se basan en modelos matemáticos modelos estadísticos simples como la regresión (regresión lineal, regresión lineal multivariante, regresión lineal no variante, regresión no lineal, se denomina regresión no lineal porque las relaciones entre los parámetros dependientes y no independientes no son lineales), estadísticos no lineales como series de potencias, Radial basis function que es una función en análisis numérico, las redes neuronales, que utiliza los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión, generalmente consisten de tres capas, de entrada, oculta y de salida.

Regresión

Una regresión es un modelo matemático para determinar el grado de dependencia en una o más variables, el conocer si existe relación entre ellas. Hay dos tipos de regresión, regresión lineal (cuando una variable independiente ejerce influencia sobre otra variable dependiente) y regresión lineal múltiple (cuando dos o más variables independientes influyen sobre una variable dependiente).

En la minería de datos, la regresión es una tarea predictiva, lo que quiere decir que predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos. El objetivo de la regresión es analizar los datos en conjunto y en base a eso, predecir lo que puede ocurrir con ese conjunto de datos en un futuro.

El análisis de regresión se basa en que al analizar la relación de las variables se identifiquen cuales tienen más impacto en el tema de interés, nos permite entender y explicar un fenómeno, prediciendo cosas a futuro, lo que nos será de ayuda para tomar decisiones yo obtener los mejores resultados.

Existen dos tipos de variables, las dependientes y las independientes, se puede trabajar con una o varias. Las dependientes es el factor más importante, el cual se está tratando de entender o predecir mientras que las independientes es el factor que se cree que puede impactar en la variable dependiente.

Clustering

En español su traducción seria "análisis de grupos", es una técnica común en el análisis de datos estadísticos y la tarea principal de la minería de datos exploratoria. El clustering, también conocido como agrupamiento, su proceso consiste en la división de los datos en grupo de objetos similares.

Las técnicas del clustering son las que utilizando algoritmos matemáticos que se encargan de agrupar objetos, algunos algoritmos son Simple k – Means, X – Means (que es una variante mejorada del k - Means), Cobweb y EM, este último es un clustering probabilístico. Los algoritmos funcionan usando la información que brindan las variables que pertenecen a cada objeto, se mide la similitud de 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 si dentro del mismo grupo, disimilar a los objetos en otros grupos. El análisis del cluster se basa que 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.

Sus aplicaciones son múltiples, el estudio de los terremotos, uso del suelo, marketing, planificación de la ciudad, y lo más importante en las aseguradoras ya que identifica los grupos de asegurados de seguros de automóviles con un alto costo promedio de reclamo.

Los métodos de agrupación son, asignación jerárquica frente a un punto, datos numéricos y/o simbólicos, determinística vs probabilística, exclusivo vs superpuesto, jerárquico vs plano, de arriba a abajo y de abajo a arriba.

Visualización

La visualización de datos nos sirve para representar gráficamente los elementos más importantes de nuestra base de datos, comparaciones de bases de datos, es la presentación de información en formato ilustrado o gráfico, se utilizan elementos visuales como cuadros, gráficos o mapas para tener una accesible de ver y comprender tendencias, valores atípicos y/o patrones en los datos.

Es importante la visualización de datos ya que ahora en la actualidad con el boom del "big data" es importante darle sentido a millones de datos, una buena visualización cuenta una historia, eliminando el ruido de los datos y resaltando la información útil.

Hay muchos tipos de la visualización de datos gráficos, infografías, cuadros de mandos (dashboards), pero la que más llama la atención son los mapas, no es lo mismo hablado una dirección con otra persona que el estar con el mapa guiándote a dónde quieres ir hasta te va hablando, guiándote.

La mayoría de los analistas de datos utilizan software avanzado para explorar y visualizar datos, que van desde hojas de cálculo sencillas con Excel o Google Sheets a softwares más avanzados de analíticas como R.

Sus aplicaciones son muy interesantes ya que después de todo un análisis les vas a mostrar historias a otras personas solo con un par de gráficos, los contadores por ejemplo al exponer su cierre de mes no van a dar sus tablas de Excel con miles de datos, si no, van a dar gráficos con sus resultados de mes, proyecciones a futuro.

Clasificación

La clasificación es una tarea predictiva, lo que quiere decir que predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos.

Consiste en el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene. Sus métodos de la clasificación análisis discriminante: 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 periódica, si se encuentra una coincidencia se agrega a los datos de clasificación, árboles de decisión: es un método analítico que a través de una representación esquemática facilita la toma de decisiones, redes neuronales artificiales: también se le conoce como sistema conexionista, es un modelo de unidades conectadas para transmitir señales.

Las características de los métodos de clasificación son precisión en la predicción, eficiencia, robustez, escalabilidad e interpretabilidad. Todas estas técnicas nos ayudan a comprender mejor la información, poder procesarla con ausencia de datos, etc.