Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Minería de Datos

Prof. Mayra Cristina Berrones Reyes

Resumen de las Técnicas de Minería de Datos

Anahí Alemán Alvarado

Matrícula: 1821952

Grupo: 03

02/10/2020

Reglas de asociación

La técnica de reglas de asociación es una búsqueda 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. Se puede aplicar para el análisis de datos de la banca, para el cross-marketing y el diseño de catálogos.

El objetivo de esta técnica es que, dado un conjunto de transacciones, se encuentren todas las reglas tomando en cuenta un umbral mínimo de soporte y un umbral mínimo de confianza. Por ejemplo, dado un conjunto de transacciones, encontrar reglas que predigan la ocurrencia de un artículo según las ocurrencias de otros artículos en la transacción. Primero, el soporte es la fracción de transacciones que contiene un itemset, el conjunto de elementos frecuentes es un conjunto de elementos con un soporte mayor o igual al umbral mínimo, recuento de soporte es la frecuencia en la que ocurre un itemset y la confianza mide que tan frecuente itemset en Y aparecen en transacciones que contienen X. Para el enfoque de fuerza bruta, es que teniendo listas todas las reglas de asociación, comprobando el soporte y la confianza, se eliminen las reglas que fallan según los umbrales.

Para las RAM enfoque 2 pasos, la generación de elementos frecuentes es en base a cuyo soporte es mayor o igual al mínimo superior. La generación de reglas es de alta confianza a partir de un conjunto de elementos frecuentes. Cada regla es una partición binaria de un conjunto de elementos frecuente.

Para las RA principio apriori, nos dice que, si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes. El soporte de un conjunto de elementos nunca excede el soporte de sus subconjuntos. Esto se conoce como la propiedad anti-monótona de soporte. Para este algoritmo, utilizamos conjuntos frecuentes (k-1) para generar candidatos a k-ítems frecuentes, además, del escaneo de la base de datos y la coincidencia de patrones para recoger los recuentos de los conjuntos de elementos candidatos. Comprimir una gran base de datos en una estructura compacta de árbol de patrones frecuentes (FP-tree), evita costosos análisis de bases de datos.

Clasificación

La clasificación es una técnica de las tareas predictivas, donde se predice el valor de un atributo basándose en los datos recolectados de otros atributos. Clasificación es el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene.

Existen algunos métodos, entre los cuales están: el análisis discriminante, que sirve para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos; árboles de decisión, que es un método a través del cual una representación esquemática facilita la toma de decisiones; reglas de clasificación, buscan términos no clasificados a manera periódica, si se encuentra coincidencia se agrega a los datos; y las redes neuronales

artificiales o sistema conexionista, es un modelo de unidades conectadas para transmitir señales.

Entre las características de estos métodos se encuentran la precisión en la predicción, la eficiencia, la robustez, la escalabilidad y la interpretabilidad.

Por ejemplo, en EUA, los profesores clasifican a los alumnos según sus notas. Si la $nota \geq 90$ pertenece al A, si $80 \leq nota < 90$ pertenece al B, si $70 \leq nota < 80$ pertenece al C, si $60 \leq nota < 70$ pertenece al D, y si $nota < 60$ pertenece al F.

Outliers

La detección de Outliers estudia el comportamiento de valores extremos que difieren del patrón general de una muestra, es decir, un valor atípico. Los valores atípicos son observaciones cuyos valores son muy diferentes a las demás observaciones del grupo de datos. Estos datos atípicos se ocasionan por errores de entrada de datos, por acontecimientos extraordinarios, por valores extremos y por otras causas no conocidas.

Los valores atípicos se calculan mediante distintos tipos de técnicas para detectarlos, estas se dividen en dos categorías, que son métodos univariantes de detección de Outliers y los métodos multivariantes de detección de Outliers. Entre las técnicas para la detección de valores atípicos están la prueba de Grubbs, de Dixon, de Tukey (diagrama de caja), el análisis de valores atípicos de Mahalanobis y la regresión simple (regresión por mínimos cuadrados).

Para la identificación de Outliers se pueden utilizar programas como R, Excel, Google Analytics, Minitab y Tableau. Ya una vez detectados los Outliers, se puede eliminar o sustituir si se corrobora que los datos atípicos se deben a un error de captura o en la medición de la variable. En caso de no deberse a un error, eliminarlo o sustituirlo ayudaría a modificar las inferencias que se realicen a partir de esa información, ya que induce a un sesgo, disminuye el tamaño de la muestra y puede afectar a la distribución y varianzas.

La minería de datos se puede aplicar para la detección de fraudes financieros, la tecnología informática y telecomunicaciones, nutrición y salud, negocios, entre otros.

Los Outliers, pueden significar error, por ejemplo, si tenemos un grupo de “edades de personas” y tenemos una persona con 160 años, seguramente sea un error de carga de datos. Puede significar límites, que son valores que se escapan del “grupo medio”, pero queremos mantener el dato modificado, para que no perjudique al aprendizaje del modelo. Puede significar punto de interés, que podrían ser los casos anómalos los que queremos detectar y que sean nuestro objetivo.

Patrones secuenciales

La 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. Las reglas de asociación secuencial expresan patrones secuenciales, esto quiere decir 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.

Las ventajas de los patrones secuenciales son la flexibilidad y la eficiencia, las desventajas son la utilización y el sesgo por primeros patrones. Algunos ejemplos de tipos de datos para estos patrones son ADN y proteínas, recorrido de clientes en un supermercado y registros de accesos a una página web. Tiene aplicación en la medicina, en el análisis de mercado y la web.

En el proceso de los patrones secuenciales $|s|$es el número de elementos en una secuencia, una k-secuencia es una secuencia con k eventos. Una subsecuencia es una secuencia que está dentro de otra, pero que cumple ciertas normas. El ítem del evento i de la subsecuencia, está dentro del evento i de la secuencia.

El método GSP, generalized sequential pattern, se divide en dos fases, Fase 1 recorre la base de datos para obtener todas las secuencias frecuentes de un elemento, y la Fase 2, Generación, que es generar k-secuencias candidatas a partir de (k-1) secuencias frecuentes; la Poda, que es podar k-secuencias candidatas que contengan alguna (k-1) secuencia no frecuente; el Conteo, obtener el soporte de las candidatas y la Eliminación, que es eliminar las k-secuencias candidatas cuyo soporte real esté por debajo del umbral de soporte mínimo de frecuencia.

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. Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo, los valores de las variables son generalmente continuos y las predicciones son usualmente sobre el futuro. Las variables pueden ser independientes, con atributos ya conocidos, o de respuesta, lo que queremos saber.

Cualquiera de las técnicas utilizadas para la clasificación y la estimación puede ser adaptada para su uso en la predicción mediante el uso de ejemplos de entrenamiento donde el valor de la variable que se predijo que ya es conocido. Los datos históricos se utilizan 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 del comportamiento futuro de los mismos.

Dentro de las aplicaciones que tiene la predicción están el revisar los historiales crediticios de los consumidores y las compras pasadas 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 a la humedad actual o predecir la puntuación de cualquier equipo durante un partido de fútbol.

Las técnicas de predicción en su mayoría se basan en modelos estadísticos simples como regresión, estadísticas no lineales como series de potencias, redes neuronales, RBF, entre otros. Entre los tipos de regresión se encuentran la regresión lineal, la regresión lineal multivariante, la regresión no lineal y la regresión no lineal multivariante.

En el Análisis de regresión simple, se pretende estudiar y explicar el comportamiento de una variable que notamos y, y que llamaremos variable dependiente o variable de interés, a partir de otra variable, que notamos x, y que llamamos variable explicativa, variable de predicción o variable independiente. Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella. La regresión no lineal es una regresión en la que las variables dependientes o de criterio se modelan como una función no lineal de los parámetros del modelo y una o más variables independientes.

Las redes neuronales, utilizan los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión. Este proceso se conoce como entrenamiento de la red neuronal. Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.

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 la regresión Lineal, que es cuando una variable independiente ejerce influencia sobre otra variable dependiente. Y la regresión Lineal Múltiple, cuando dos o más variables independientes influyen sobre una variable dependiente.

En minería de datos, es parte de la categoría Predictivo y 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 nos permite explicar un fenómeno y predecir cosas acerca del futuro, por lo que es de ayuda para tomar decisiones y obtener los mejores resultados. Permite examinar la relación entre dos o más variables e identificar cuáles son las que tienen mayor impacto en un tema de interés. La variable independiente es el factor más importante y la variable independiente es el factor que uno cree que puede impactar en nuestra variable dependiente.

La idea de la regresión lineal consiste en obtener una ecuación de la forma $y = mx + b$ que se ajuste más a los datos, donde m es la pendiente de los datos y $b = \bar{y} - m\bar{x}$.

Para determinar qué tan bueno es el ajuste, existen diferentes parámetros estadísticos, pero en este caso se utiliza el coeficiente de determinación $R = \frac{\sigma xy}{\sigma x\, \sigma y}$.

Visualización

La visualización de datos es la presentación de información en formato ilustrado o gráfico. Al utilizar elementos visuales como cuadros, gráficos o mapas, nos proporciona una manera accesible de ver y comprender tendencias, valores atípicos y/o patrones en los datos.

Existen diferentes tipos de Visualización de datos y cada tipo de elemento visual se debe utilizar para representar la información de la mejor forma. Entre los más comunes están: los gráficos, que es el tipo más común y conocido, utilizados para representar datos de manera sencilla, como Gráficos Circulares, Líneas, Columnas y Barras aisladas o agrupadas, Burbujas, áreas, Diagramas de Dispersión y Mapas de tipo Árbol. Los mapas, la visualización de datos en mapas para conocer, por ejemplo, la localización de una flota de vehículos en tiempo real o bien la de las tiendas de un supermercado o los cajeros automáticos del banco en un mapa. Infografías, que son colecciones de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente, ayudan a procesar más fácil la información compleja. Los cuadros de Mando, que son una herramienta que permite saber en todo momento el estado de los indicadores del negocio: de ventas, económicos, de producción, de recursos humanos, etc. y que nos dice lo que está pasando en la empresa para poder tomar decisiones adecuadas, ya sean correctivas o de planeación.

La mayoría de los analistas de datos utilizan software avanzado para explorar y visualizar datos. Y las herramientas de software van desde Hojas de Cálculo sencillas con Excel o Google Sheets a software de analítica más sofisticado, como R.

Tiene aplicación para comprender la información con rapidez, identificar relaciones y patrones, identificar tendencias emergentes, comunicar la historia a otras personas. A medida que la "era del big data" entra en pleno apogeo, la visualización es una herramienta cada vez más importante para darle sentido a los billones de filas de datos que se generan cada día. La visualización de datos ayuda a contar historias seleccionando los datos en una forma más fácil de entender, destacando las tendencias y los valores atípicos.

Clustering

El Clustering o Agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares. 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 y disimilar a los objetos en otros grupos. El análisis de cluster, es 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.

Entre las aplicaciones están las áreas de aseguradoras, en la identificación de grupos de asegurados de seguros de automóviles con un alto costo promedio de reclamo. El uso del suelo, en la identificación de áreas de uso similar de la tierra en una base de datos de observación de la tierra. En Marketing, ayudar a los profesionales de marketing a descubrir distintos grupos en sus bases de clientes. En la planificación de la ciudad, identificación de grupos de casas según su tipo de casa, valor, y ubicación geográfica. En estudios de terremotos, los epicentros de un terremoto deben agruparse a lo largo de fallas continentales.

Los métodos de agrupación se clasifican por asignación jerárquica frente a punto, datos numéricos y/o simbólicos, determinista vs probabilística, exclusivo vs superpuesto, jerárquico vs plano, de arriba abajo y viceversa.

Existen diversos algoritmos de Clustering, entre los más conocidos están el Simple K-Means, este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar. El X-Means, este algoritmo es capaz de obtener en el rango de k-min y k-max el número óptimo de clusters, dando de esta manera más flexibilidad al usuario. Cobweb, pertenece a la familia de algoritmos jerárquicos y se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia. El EM, pertenece a una familia de modelos que se conocen como Finite Mixture Models, los cuales se pueden utilizar para segmentar conjuntos de datos, está clasificado como un método de particionado y recolocación, o sea, Clustering Probabilístico.