UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

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

Nombre: Manuel de Jesús Vázquez Bocanegra

Matricula:1823593

Reglas de Asociación

Las reglas de asociación es una técnica que se utiliza en la inteligencia artificial en el data meeting lo que hace es describir una regla como su nombre lo indica de asociación entre los conjuntos de datos relevantes. Es la búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de elementos u objetos en bases de datos de transacciones, relacionales y otros repositorios de información disponible.

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 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

El tema de clasificación consiste básicamente en la predicción de variables cualitativas, esto es, dada una observación tú quieres poder predecir si va a pretender a una clase especifica o no (o incluso la probabilidad de que pertenezca).

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.

Detección de 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.

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.

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.

En esta técnica existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo, los valores son generalmente continuos y como se mencionó anteriormente, las predicciones son a menudo sobre el futuro.

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.

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.

Visualización de datos

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.

La visualización de datos es más importante 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 datos que se generan cada día h ayudar a contar 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.