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

Resumen Técnicas

Arturo Del Ángel De La Cruz 1809895

2 de Octubre de 2020

Reglas de Asociación

Las reglas de asociación son derivadas de un análisis que extrae información por coincidencias, con el objetivo de encontrar relaciones dentro de un conjunto de transacciones. Estas nos permiten encontrar combinaciones que ocurren con más frecuencia en una base de datos y medir la fuerza e importancia de estas combinaciones. Dentro de sus muchas aplicaciones tenemos, definir patrones de navegación, promociones de pares de productos, soporte para la toma de decisiones, análisis de información de ventas, distribución de mercancía, segmentación de clientes. Entre sus tipos tenemos la asociación cuantitativa la cual cuenta con la asociación booleana la cual hace asociaciones entre la presencia ausencia de un ítem y la asociación cuantitativa describe asociaciones entre ítems cuantitativos o atributos. Otro tipo es la asociación multidimensional donde tenemos la asociación unidimensional y la multidimensional, donde la primera asocia si los ítems o atributos de la regla se referencian en una sola dimensión y la segunda si los ítems o atributos de la regla se referencian en dos o más dimensiones. Por último tenemos las asociaciones multinivel donde los ítems referenciados a un solo nivel de abstracción son de una asociación de un nivel y los ítems relacionados a varios niveles son de una asociación multinivel. Dentro de las métricas de interés tenemos el soporte, el cual es el número de veces o frecuencia con que 2 ítems aparecen juntos en una base de datos de transacciones, la confianza que se obtiene con el soporte y mide la fortaleza de la regla y finalmente está el lift el cual refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente.

Outliers

Los outliers en la minería es la detección de datos anómalos o comportamientos inusuales en los datos, su definición formal sería la observación que se desvía mucho del resto de las observaciones apareciendo como una observación sospechosa que pudo ser generada por mecanismos diferentes al resto de los datos, esta técnica se puede aplicar en el aseguramiento de ingresos en las telecomunicaciones, detección de fraudes financieros, seguridad y detección de fallas, pero ¿cómo aplicamos esto?, para esto se realizan unas pruebas estadísticas no paramétricas para comparar los datos y su comportamiento basados en la capacidad del algoritmo, una de estas pruebas puede ser la regresión donde nosotros obtenemos una serie de datos y al final estos se ven reflejados en una gráfica para tener una mejor visualización de estos, de esta forma podemos ver como algunos de los puntos seguirán una tendencia pero podemos tener los outliers que van a estar más separados de la tendencia y de los otros datos lo que va a causar un cambio que puede ser significativo en la tendencia y aquí es donde nosotros podemos identificar que hay unos

datos que son diferentes a los demás, de ahí realizar las aplicaciones y poder verificar si hay algún problema para poder solucionarlo.

Regresión

La regresión lleva muchos años avanzando, vemos como desde 1805 se crea el primer modelo hasta los modelos actuales los cuales ya son computarizados. La regresión es una técnica de minería de datos la categoría predictiva. Esta predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos. Se encarga de analizar un vínculo entre una variable que es dependiente y otra u otras que serían independientes por lo que nos encontramos con dos regresiones, la regresión lineal simple y la regresión lineal múltiple. Bajo estas dos regresiones la regresión lineal simple solo trata de una variable regresora, contamos con dos betas las cuales serán nuestras constantes en esta regresión, beta cero será la constante sin una variable y beta uno será el coeficiente que hace producto con la variable regresora, la estimación de la respuesta debe ser una recta que proporcione un buen ajuste a los datos observados. En cambio la regresión lineal múltiple se dice que es lineal porque la ecuación modelo es una función lineal de los parámetros desconocidos, esta contiene k regresores por lo que va a contener k más un regresores donde tenemos la beta cero y las demás betas para los k regresores. Estas regresiones tienen muchas aplicaciones como en la medicina, informática, industria, estadística, etc.

Predicción

Para hacer un buen modelo de predicción hay unos elementos previos que debemos considerar los cuales son: Definir adecuadamente nuestro problema, Recopilar datos, Elegir una medida o indicador de éxito y preparar los datos. Ya teniendo los elementos recién comentados debemos dividir los datos donde el 70% será el conjunto de entrenamiento, 15% el conjunto de validación y otro 15% para el conjunto de pruebas. Los árboles de decisión son un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Los árboles de decisión se dividen en dos tipos, árboles de regresión en los cuales la variable de respuesta es cuantitativa y árboles de clasificación en los cuales la variable de respuesta es cualitativa. Dentro de la estructura de un árbol de decisión encontraremos 3 tipos de nodos, el primer nodo o también llamado nodo raíz es donde se hace la primera división en función de la variable más importante, los nodos internos o intermedios que se da después del primer nodo y vuelven a dividir el conjunto de datos en función de las variables, por último tenemos los nodos terminales u hojas los cuales se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva. La información de cada nodo es: Condición, Gini, Samples, Value y Class. Un conjunto de árboles de decisión conforman un Bosque

aleatorio, el cual es una técnica de aprendizaje la cual obtiene un mejor rendimiento de generalización. Una validación cruzada se emplea para estimar el test error rate de un modelo y así evaluar su capacidad predictiva, a este modelo se le conoce como model assessment.

Clustering

Técnica que consiste en agrupar puntos de datos y teniendo estos grupos realizar particiones basándose en las similitudes que lleguen a presentar estos datos. Algunos usos que puede traer esta técnica son para la investigación del mercado, identificar comunidades, prevenir el crimen, procesamiento de imágenes, entre otros. Hay que considerar que los datos en caso de necesitarlo sean transformados, si son variables cuantitativas deben tener las mismas mediciones, en variables binarias deben seguir con las características de los datos binarios y en variables categóricas podemos realizar una binarización. Tenemos 4 tipos de análisis básicos, primero está el Centroid Based Clustering donde cada cluster es representado por un centroide y su algoritmo más utilizado es el de k-medias, está también el Connectivity Based Clustering el cual consiste en agrupar los datos similares, en este tipo un cluster contiene a otros clusters y su algoritmo más usado es el Hierarchical clustering, el tipo de Distribution Based Clustering cada cluster pertenece a una distribución normal, se dividen con base a la probabilidad, su algoritmo es Gaussian mixture models, por último tenemos el Density Based Clustering donde los clusters se definen por áreas de concentración y conecta puntos con distancia relativamente pequeña. En el método de las k-medias la k representa el número de clusters, la varianza de los clusters disminuye al aumentar k. La mejor k la podemos seleccionar con el método del codo donde graficamos la varianza con respecto a las k y donde haya un menor cambio significativo se tomará esa k

Visualización

La visualización de los datos es la representación gráfica de la información y de los datos. Utiliza elementos visuales como cuadros, gráficos y mapas las cuales son herramientas de visualización de datos que nos ayudan a ver de manera más fácil y comprender las tendencias, valores atípicos y patrones que llegue a haber en los datos, esta técnica es esencial para analizar una gran cantidad de información y tomar buenas decisiones basadas en los datos. Tenemos algunos tipos de visualizaciones uno de ellos es el de elementos básicos de representación de datos que como ya se comentó antes ayudan a ver mejor los datos y entre estos están las gráficas de cualquier tipo como las de barras, columnas, puntos, etc., los mapas tanto como de burbujas, mapas de calor, entre otros y las tablas de diferentes tipos, otro tipo de visualización son los cuadros de mando el cual es una composición compleja de visualizaciones individuales que guardan una coherencia y una

relación temática entre ellas, por último tenemos el tipo que es Infografías las cuales no están destinadas al análisis de variables si no a la construcción de narrativas a partir de los datos, por lo que se usan para contar "historias". Hay varios estándares web que se han desarrollado en los últimos años para la evolución de aplicaciones web donde se encuentran unos ejemplos como HTML5, CSS3, SCV Y WebGL. Es importante una visualización de datos ya que el mundo está avanzando para convertirse en un mundo basado en datos.

Patrones Secuenciales

Los patrones secuenciales se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias, se describe como el modelo de compras que hace un cliente particularmente o un grupo de clientes relacionando las distintas transacciones efectuadas por ellos a lo largo del tiempo por lo que los eventos se enlazan con el tiempo. Hay que buscar asociaciones entre un evento y el tiempo, el objetivo es descubrir las relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos, los patrones secuenciales utilizan reglas de asociación secuenciales donde se expresan patrones de comportamiento secuencial. Dentro de sus características podemos encontrar que el orden importa, su objetivo es encontrar patrones de secuencia, tener en cuenta que una serie es una lista ordenada de itemsets, donde cada itemset es un elemento, el tamaño de la secuencia es el número de elementos, la longitud es la cantidad de ítems, el soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S y las secuencias frecuentes son las subsecuencias de una secuencia que tienen un soporte mínimo. Para la resolución de problemas necesitamos agrupamiento de patrones secuenciales donde separamos en grupos los datos, clasificación con datos secuenciales los cuales se expresan como patrones de comportamiento secuenciales, reglas de asociación con datos secuenciales las cuales se presentan cuando los datos contiguos presentan algún tipo de relación.

Clasificación

Esta técnica de la minería de datos es la que se usa más comúnmente, esta organiza o mapea un conjunto de atributos por clase dependiendo de sus características. Una de las cosas para la que esta técnica puede funcionar es entrenar un modelo usando los datos que se recolectan para hacer predicciones futuras. Hay varias técnicas de clasificación donde está la clasificación Bayesiana donde tenemos una hipótesis H para una evidencia E, representaremos aquí como p(A) la probabilidad del suceso y como p(A|B) la probabilidad del suceso A condicionada por el suceso B, las redes neuronales es otra técnica de clasificación, esta trabaja directamente con números y en caso de que se deseé trabajar con datos nominales, estos deberán de enumerarse, consiste de tres fases las cuales son de entrada, oculta y de salida, estos se pueden ver como una gráfica dirigida, un árbol de

decisión es otra técnica en la cual hay una serie de condiciones organizadas en forma jerárquica, a modo de árbol, se utilizan para problemas con datos numéricos y categóricos, útiles en clasificación, agrupamiento y regresión, también se encuentran las técnicas de clasificación basada en asociaciones y Support Vector Machines (SVM).