Universidad autónoma de Nuevo León

Minería de datos

Resúmenes Técnicas

Profesora: Mayra Cristina Berrones Reyes

Alumno: Miguel Eduardo Ovalle Blanco 1801990

Grupo: 002

2 de octubre del 2020

Técnicas de minería de datos

Reglas de asociación

Estas reglas se derivan de análisis que extrae información por coincidencias, con el objetivo de encontrar relaciones en atributos que ocurren de forma consecutiva. Esto quiere decir que si ocurre un evento por consecuente tenemos otro, $A => B$. Esto nos ayuda en que encontremos combinaciones en bases de datos transaccionales y medir la importancia de estas.

Existen diferentes tipos de reglas de asociación dependiendo de diferentes aspectos: si es con el tipo de valores (Asociación cuantitativa) las cuales si es por ausencia/presencia (Booleana) o entre ítems cuantitativos o atributos (Cuantitativa), también tenemos que dependen de las dimensiones de los datos (Asociación multidimensional) que si es de una dimensión (Unidimensional) o si los ítems tienen más dimensiones (Multidimensional) y por último tenemos por medio de niveles de abstracción (Asociación multinivel) que si son de un solo nivel (De un nivel) o de más niveles (Multinivel).

Entre las métricas de interés para esto tenemos que esta el soporte es el número de veces o la frecuencia (relativa) con que A y B aparecen juntos con una base de datos de transacciones

$$Soporte(A => B) = P(A \cap B)$$
$$= \frac{Frecuencia\ en\ que\ A \cap B\ aparecen\ en\ las\ transacciones}{Total\ de\ transacciones}$$

Para esto se debe de tener un soporte mínimo. Otra de las métricas es la confianza la cual nos dice parte de la relación entre dos ítems, que tan fuerte es su relación, y esta expresada en el cociente de la regla entre el antecesor

$$Confianza(A => B) = \frac{Soporte(A => B)}{Soporte(A)} = \frac{P(A \cap B)}{P(A)}$$

Y por último tenemos el lift, que indica aumento de la probabilidad de consecuente cuando ocurre el antecesor, si el valor es mayor a 1 es una relación fuerte, si tiende a 1 es una relación al azar y si es menor a 1 es una relación débil.

$$Lift(A => B) = \frac{Soporte(A => B)}{Soporte(A) * Soporte(B)} = \frac{P(A \cap B)}{P(A) * P(B)}$$

Outliers

La calidad de los datos que manejan las organizaciones es de gran importancia a la hora de ser analizados para la obtención de información que permita la toma de decisiones empresariales, datos con errores o problemas pueden conducir a obtener información imprecisa y a la vez a tomar decisiones erróneas. Entre los posibles problemas que pueden

presentar los datos, se encuentran los conocidos como valores atípicos o "Outliers". Un valor atípico u Outlier se define como una observación de datos que es muy diferente del resto de los datos observados de una medida específica. A tal punto que a menudo contiene información útil sobre el comportamiento anormal del sistema descrito.

Se pueden clasificar en Univariantes (análisis de una única característica o cualidad de un conjunto de datos) mediante la fórmula

$$\frac{|x\bar{x}|}{s} > k$$

Donde x es una observación, $\bar{x}$ es la media de los datos y s la desviación estándar de los mismos, el valor de k esta entre 2 o 3. También hay Los valores atípicos multivariantes son observaciones que se consideran extraños no por el valor que toman en una determinada variable, sino en el conjunto de aquellas.

También hay otras formas de detectar valores atípicos como por ejemplo utilizando clustering gracias a los algoritmos K-means y PAM. También se pueden calcular basados en la distancia, y de igual manera por la densidad que tiene ese punto y de otra manera se pueden combinar los dos métodos con lo cual se tendría mejor precisión en la localización de esos puntos con datos atípicos.

Regresión

La regresión es una técnica de minería de datos de la categoría predictiva. Predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos y se encarga de analizar el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática.

Si sólo se trata de una variable regresora, se llama regresión lineal simple y tiene la ecuación $y = \beta_1 x + \beta_0 + e$, e es una variable normalmente distribuida con $E(e) = 0\ y\ Var(e) = \sigma^2$. Para obtener la estimación de la recta se usa el modelo ajustado por mínimos cuadrados

$$\hat{\beta}_0 = \bar{y} + \hat{\beta}_1 x \qquad\qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

Un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos. En general, se puede relacionar la respuesta "y" con los k regresores, o variables predictivas bajo el modelo: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$

De tal manera que la función de mínimos cuadrados queda como:

$$S(\beta_0, \beta_1, .., \beta_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

Donde se debe minimizar en donde de forma matricial

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{12} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Donde el estimador de mínimos cuadrados es $\hat{\beta} = (X'X)^{-1}X'y$, siempre y cuando exista $(X'X)^{-1}$

Entre algunas aplicaciones están que las inmobiliarias usan las regresiones lineales para predecir el valor de un inmueble con las variables de metros cuadrados, relación de baños por dormitorios, año de construcción y código postal.

Predicción

Para esto tenemos que fijar bien el objetivo de nuestro problema, recopilar datos, elegir un indicador de éxito y preparar dichos datos de los cuales el 70% serán de entrenamiento (que es el modelo reconozca lo que debe de resultar), 15% en un conjunto de validación y el otro 15% en conjunto de pruebas.

Tenemos los árboles de decisión que son un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Para dividir el espacio muestral en subregiones es preciso aplicar una serie de reglas o decisiones, para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones.

Si una subregión contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en subregiones menores que integran datos de la misma clase. Los árboles se pueden clasificar en dos tipos que son: 1. Árboles de regresión en los cuales la variable respuesta y es cuantitativa. 2. Árboles de clasificación en los cuales la variable respuesta y es cualitativa

Dentro de un árbol de decisión existe: *Primer nodo o nodo raíz*, en él se produce la primera división en función de la variable más importante; Nodos internos o intermedios, tras la primera división, vuelven a dividir el conjunto de datos en función de las variables; Nodos terminales u hojas que se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva

Podemos clasificar los nodos como: Nodos de decisión que tienen una condición al principio y tienen más nodos debajo de ellos y Nodos de predicción que no tienen ninguna condición ni nodos debajo de ellos.

La información de cada nodo es la siguiente: Condición, Si es un nodo donde se toma alguna decisión, Gini: Es una medida de impureza que nos dice que tan mezclada esta con otras clases $gini = 1 - \sum_{k=1}^{n} p_c^2$, Samples es el número de muestras que satisfacen las

condiciones necesarias para llegar a este nodo, Value es la cantidad de muestras de cada clase llegan a este nodo y Class es la clase que se les asigna a las muestras que llegan a este nodo.

También tenemos a los random forest que son una técnica de aprendizaje automático supervisada basada en árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar. Esta mejora en la generalización la consigue compensando los errores de las predicciones de los distintos árboles de decisión. Para asegurarnos que los árboles sean distintos, lo que hacemos es que cada uno se entrena con una muestra aleatoria de los datos de entrenamiento. Esta estrategia se denomina bagging (consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados)

Al realizar todo esto también tenemos lo que se denomina validación cruzada, Se emplea para estimar el test error rate de un modelo y así evaluar su capacidad predictiva, a este proceso se le conoce como model assessment. También se puede emplear para seleccionar el nivel de flexibilidad adecuado.

Clustering

El Clustering consiste en agrupar un conjunto de objetos en subconjuntos de objetos llamados Clusters donde cada Cluster está formado por una colección de objetos que considerados similares entre sí, pero que son distintos respecto a los objetos de otros Clusters. Al ser una técnica de aprendizaje no supervisada no tiene una clase de respuesta. Es decir, no se cuenta con información sobre la estructura del dominio de salida por lo que después de agrupar observaciones es necesario asociarle un significado o característica distintiva a cada clúster.

Algunos ejemplos del uso del clustering son: Investigación de mercado, identificar comunidades en redes sociales, prevención de crímenes y procesamiento de imágenes. Para usar este modelo necesitamos transformar los datos así para **variables cuantitativas** se recomienda una transformación si estas presentan diversas unidades de medida siendo la más popular la estandarización, en **variables binarias** no suelen sufrir transformaciones y las **variables categóricas** son convertidas en variables numéricas por binarización (presencia/ausencia)

Existen diferentes diferentes análisis de clusters:

Centroid Based Clustering: Cada clúster es representado por un centroide. Los clusters se construyen basados en la distancia de punto de los datos hasta el centroide. El algoritmo más usado de este tipo es el de *K-medias.*

Connectivity Based Clustering: Los clusters se definen agrupando a los datos más similares o cercanos basándonos en la premisa de que los puntos más cercanos están más relacionados que otros puntos más lejanos. Puede comenzar de un cluster grande a

dividirse en pequeños o viceversa. *Hierarchical clustering* es un algoritmo de clustering perteneciente a este tipo.

Distribution Based Clustering: En este método cada cluster pertenece a una distribución normal, la idea es que los puntos son divididos basados en la probabilidad de pertenecer a la misma distribución normal. Un algoritmo usado en este caso es *Gaussian mixture models*.

Density Based Clustering: Los clusters son definidos por áreas de concentración. Este método comienza buscando áreas de puntos concentrados (cercanos) y asigna esas áreas al mismo cluster. Se trata de conectar puntos cuya distancia entre sí es considerada pequeña.

Pasos para el algoritmo de k medias

1. Centroides: Elegimos k datos aleatorios que pasarán a ser los centroides representativos de cada cluster
2. Distancias: Analizamos la distancia de cada dato al centroide más cercano, perteneciendo a su cluster.
3. Media: Obtener media de cada cluster y este será el nuevo centro.
4. Iterar: Repetimos el proceso hasta que los clusters no cambien

La varianza al ser una parte de este proceso se tiene que se reduce al aumentar la k, ya que es el numero de las varianzas de los cluster. Con esto una manera efectiva de encontrar el número indicado es el método del codo que consiste en graficar la reducción de la varianza total a medida que k aumenta y en un punto la reducción de la varianza no disminuirá de forma significativa entre un valor k y otro. Con lo cual ese punto se seleccionará.

Visualización de Datos

Se define como la representación gráfica de información y datos. Al utilizar elementos visuales como cuadros, gráficos y mapas, las herramientas de visualización de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos.

Algunos tipos de visualización son:

1. Elementos básicos de representación de datos: En estos se encuentran visualizaciones básicas como gráficas (barras, líneas, columnas, puntos, “tree maps”, tarta, semi-tarta, etc.), mapas (burbujas, coropletas (o mapa temático), mapa de calor, de agregación (o análisis de drilldown)) y tablas (con anidación, dinámicas, de drilldown, de transiciones, etc.)
2. Cuadro de mando. Es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas

3. Infografías. Esta narrativa se construye a través de la disposición de la información en la que las visualizaciones se combinan con otros elementos como: símbolos, leyendas, dibujos, imágenes sintéticas, etc.

Los conjuntos de habilidades están cambiando para adaptarse a un mundo basado en los datos. Para los profesionales es cada vez más valioso poder usar los datos para tomar decisiones y usar elementos visuales para contar historias con los datos para informar quién, qué, cuándo, dónde y cómo. La visualización de datos se encuentra justo en el centro del análisis y la narración visual.

Patrones Secuenciales

Se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias. Es una clase especial de dependencia en las que el orden de acontecimientos es considerado y describe 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.

Entre las características de los patrones se encuentran: El orden importa, su objetivo es encontrar patrones en secuencia, una secuencia es una lista ordenada de itemsets, donde cada itemsetes un elemento de la secuencia; el tamaño de una secuencia es su cantidad de elementos (itemsets), la longitud de una secuencia es su 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 (o patrones secuenciales) son las subsecuencias de una secuencia que tienen un soporte mínimo.

Para la resolución de estos problemas tenemos:

- Agrupación de Patrones Secuenciales: Se define como la tarea de separar en grupos a los datos, de manera que los miembros de un grupo sean muy similares entre sí, y al mismo tiempo sean diferentes a los objetivos de otros grupos.

  Para la creación de agrupamientos:
  - Se selecciona arbitrariamente el centro del primer agrupamiento.
  - Posteriormente, se procesan secuencialmente los demás patrones mediante cálculos de distancia.

  Cada M patrones se mezclan agrupamientos, estos pueden ser:
  - Mezcla por cercanía.
  - Mezcla por tamaño.
  - Mezcla forzada.

- Clasificación de datos secuenciales: Éstos expresan patrones de comportamiento secuenciales, es decir que se dan en instantes distintos (pero cercanos) en el tiempo.
- Reglas de asociación con datos secuenciales: Se presenta cuando los datos contiguos presentan algún tipo de relación.

Clasificación

La clasificación es la técnica de minería de datos más comúnmente aplicada, que organiza o mapea un conjunto de atributos por clase dependiendo de sus características. Con ello se entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras.

Primeramente, para aplicar las técnicas de clasificación se necesita: limpieza de los datos (tratamiento del ruido y de valores faltantes), Análisis de relevancia (algunos atributos en los datos pueden ser irrelevantes o redundantes. Eliminar dichos atributos mejora la eficiencia y la eficacia) y Transformación de datos (se pueden hacer generalizaciones de los datos a conceptos de mayor nivel. También se pueden normalizar los datos).

Después tenemos diferentes maneras de evaluar los métodos como sería: Precisión en la predicción (capacidad de predecir correctamente), Eficiencia (Costos computacionales), Robustez (Habilidad para funcionar con ruido y ausencia de ciertos valores), Escalabilidad (Habilidad para trabajar con grandes cantidades de datos) e Interpretabilidad (Entendimiento y comprensión que brinda)

Existen diferentes técnicas de clasificación como lo son:

- Clasificación por inducción de árbol de decisión
- Clasificación Bayesiana
- Redes neuronales
- Support Vector Machines (SVM)
- Clasificación basada en asociaciones