RESUMÉN DE LAS TÉCNICAS DE MINERIA DE DATOS

1. Reglas de Asociación

Es un tipo de análisis que extrae información por coincidencias, con el objetivo de mostrar relaciones en conjuntos de datos que tiendan a ocurrir en forma conjunta.

Se definen como: A (Antecedente) ⟹ B (Consecuencia)

Las reglas de asociación nos permiten: encontrar combinaciones de artículos o ítems que ocurren con mayor frecuencia en bases de datos y miden la fuerza e importancia de estas combinaciones.

Esta técnica se puede aplicar al definir patrones de navegación dentro de una tienda, en promociones de pares de productos por ejemplo las hamburguesas y la Cátsup, en soporte para la toma de decisiones. También en el análisis de información de ventas, distribución de mercancías en tiendas y en la segmentación de clientes con base en patrones de compra.

Existen tres tipos de asociación que son: Cuantitativa, Multidimensional y Multinivel.

La asociación cuantitativa se divide en booleana que son asociaciones entre la presencia o ausencia de un ítem y la cuantitativa que describe las asociaciones entre ítems cuantitativos o atributos. La asociación multidimensional se divide según la dimensión de los datos involucrando la unidimensional o multidimensional. Con base en los niveles de abstracción la asociación multinivel se divide en la de un nivel y la multinivel.

En las métricas de interés están el soporte, la confianza y LIFT. El soporte se define como el número de veces o la frecuencia con que A y B aparecen juntos en la base de datos de transacciones y se expresa como: Soporte (A => B) = P (A∩B) en donde la fórmula es: $\frac{\text{Frecuencia en que A} \cap \text{B aparecen en las transacciones}}{\text{Total de transacciones}}$. La confianza se define como el cociente del soporte de la regla y el soporte del antecedente solamente. Finalmente el LIFT refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente, donde: $\text{Lift A} \Rightarrow \text{B} = \frac{\text{Soporte (A=>B)}}{\text{Soporte A} * \text{Soporte(B)}}$ y tenemos tres casos; si el Lift es mayor a 1 representa una relación fuerte y una frecuencia mayor que el azar, si es igual a 1 representa una relación al azar y si es menor a 1 representa una relación débil y una frecuencia menor que el azar.

2. Outliers

En español a esta técnica se le llama como datos atípicos que es la detección de datos raros o comportamientos inusuales en los datos. Este dato atípico es la observación que se desvía mucho del resto de las observaciones y apareciendo de una manera sospechosa que puede ser generada por mecanismos diferentes al resto de los datos.

Las aplicaciones de esta técnica son principalmente en el aseguramiento de ingresos en las telecomunicaciones, detección de fraudes financieros y en la seguridad y detección de fallas.

Para este método se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos.

La mayoría de los trabajos existentes sobre la detección de outliers yacen en el campo de la estadística. Existen muchas maneras de detectarlos y estos métodos se han diseñado para diferentes circunstancias que son: la distribución de los datos, si los parámetros de distribución son conocidos o no, el número de outliers esperados y el tipo de outliers esperados.

3. Regresión

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

Tenemos dos tipos de regresión que son: Regresión Lineal Simple y Regresión Lineal Múltiple.

En la regresión simple tenemos una sola variable regresora y esta tiene como modelo $y = \beta_o + \beta_1 x + e$ donde "e" es una variable aleatoria normalmente distribuida y que tiene una media igual a cero y una varianza de σ^2.

Una forma para estimar la regresión lineal es la estimación por mínimos cuadrados en donde el modelo ajustado que se utiliza es: $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} x$ y $\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$. Ahora hablando de la regresión lineal múltiple tenemos la variable respuesta "y" con k regresores o más de una variable predictiva bajo el modelo: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$. La regresión lineal tiene muchas

aplicaciones en la vida real como en la medicina, la informática, comportamiento humano, en la industria y en estadística.

4. Partición

Los elementos que se necesitan para hacer un buen modelo de predicción son: definir adecuadamente nuestro problema, recopilar los datos, elegir una medida o indicador de éxito y finalmente prepara los datos.

La manera de dividir los datos es de 70%, 15% y 15% al conjunto de entrenamiento, de validación y de pruebas respectivamente.

El árbol de decisión es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta. Para esto se aplica una serie de reglas o decisiones donde cada subregión debe contener la mayor proporción posible de individuos de una de las poblaciones. Los arboles de decisión se pueden clasificar en dos: 1. variable de respuesta y cualitativa. 2. Variable de respuesta y cualitativa. La estructura básica de un árbol de decisión está compuesta por: nodos que leen de arriba para abajo. Hay tres tipos de nodos que son: primer nodo o nodo raíz, nodos internos o intermedios y los nodos terminales u hojas.

El árbol de clasificación consiste en hacer preguntas del tipo ¿$xk \leq c$? para las covariables cuantitativas o preguntas del tipo ¿$xk = nivelj$? para las covariables cualitativas, de esta forma el espacio de las covariables es divido en hiperrectángulos y todas las observaciones que queden dentro de un hiperrectángulo tendrán el mismo valor grupo estimado.

Gini es una medida de impureza. Si esta vale 0 significa que el nodo es totalmente puro. Si las clases de cada nodo están muy mezcladas se considera como impureza. Para calcular la impureza Gini, se usa la formula: $gini = 1 - \sum_{k=1}^{n} p_c^2$

El árbol de regresión consiste en hacer preguntas de tipo ¿$xk \leq c$? para cada una de las covariables, de esta forma el espacio de las covariables es divido en hiperrectángulos y todas las observaciones que queden dentro de un híper-rectángulo tendrán el mismo valor estimado $\hat{y}$. Tiene como pasos los siguientes: 1.Encontrar la covariable que permita predecir la mejor variable respuesta. 2 Encontrar el punto de corte sobre esa covariable. 3 Repetir los pasos anteriores hasta que se alcance el criterio de parada. Tiene como ventajas que son fáciles de entender e interpretar, sencillos, las covariables pueden ser cualitativas o cuantitativas y que no se exigen los supuestos distribucionales.

Random Forest es 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 estrategia se denomina bagging. Bagging consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.

Los random forest tienen como ventaja que pueden aprender cualquier correspondencia entre datos de entrada y resultado a predecir y como desventaja que no son buenos extrapolando porque no siguen un modelo conocido

La validación cruzada se emplea para se emplea para estimar el test error rate de un modelo y evaluar su capacidad predictiva, a este proceso se le conoce como model assessment

5. Clustering

Es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear particiones basándonos en similitudes.

El clustering se puede usar para investigación de mercado, identificar comunidades, prevención de crimen y procesamiento de imágenes, entre otras cosas. Transforma datos con variables cuantitativas, binarias y categóricas.

Los tipos básicos del clustering son: 1.Centroid Based Clustering, 2.Connectivity Based Clustering, 3.Distribution Based Clustering y 4.Density Based Clustering. En el primero cada cluster se representa por un centroide y se realizan varias iteraciones hasta llegar al mejor resultado, el algoritmo más usado es el de k-medias. 2. Estos se definen agrupando datos similares y cercanos, tiene como característica principal que un cluster contiene otros clusters, el algoritmo que más se usa es el de Hierarchical clustering. 3. Cada cluster pertenece a una distribución normal, un algoritmo que pertenece a este tipo es Gaussian Mixture Models. 4. En este último los clusters se definen por áreas de concentración conectando puntos con distancia pequeña.

El método de las K-medias es un algoritmo que se basa en los centroides donde k es el número de clusters y se definen por el usuario, eligiendo k datos aleatorios que pasarán a ser los centroides representativos de cada cluster, se analiza la distancia de cada dato al centroide más cercano, perteneciendo a su cluster, obtenemos la media de cada cluster y este será el nuevo centro para finalmente repetir el proceso hasta que los clusters no cambien.

6. Visualización de datos

Es la representación grafica de información y datos utilizando elementos visuales, estos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. Este método es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos.+

Hablaremos de tres diferentes tipos de visualización que son: los elementos básicos de representación de datos, cuadros de mando e infografías. Los básicos son graficas, mapas y tablas. Los cuadros de mando son una composición de visualizaciones individuales que guardan una cierta relación temática, estos son utilizados principalmente en organizaciones para el análisis de conjuntos de variables y para la toma de decisiones. Las infografías son para la construcción de narrativas a partir de datos, estas se utilizan para contar "historias", las narrativas no se construyen a través de texto sino con la información en la que las visualizaciones se combinan como: símbolos, leyendas, dibujos, etc...

Los software de visualización de datos son: HTML5, CSS3, SCV y WebGL. La visualización de datos es importante en cualquier empleo porque para un profesional cada vez es más valioso poder usar datos para tomar decisiones y usar los elementos visuales para contar e informar quién, qué, cuándo, dónde y cómo.

7. Patrones Secuenciales

Este método se especializa en analizar datos y encontrar subsecuencias dentro de un grupo de secuencias. Describe el modelo de compras que un cliente particularmente o un grupo de clientes donde se relacionan las distintas transacciones efectuadas a lo largo del tiempo. Estos eventos enlazan con el paso del tiempo.

Para este método se buscan asociaciones de la forma "si sucede el evento X en el instante de tiempo t entonces sucederá el evento Y en el instante t+n".

Las características de los patrones secuenciales son: el orden importa, el objetivo es encontrar secuencias con patrones, las secuencias son una lista ordenada de ítems y cada ítem es un elemento de la secuencia. El tamaño de la secuencia es la cantidad de ítemsets y la longitud la cantidad de ítems.

El método es utilizado en la medicina, biología y bioingeniería, web, análisis de mercado, la distribución y comercio, aplicaciones financieras y de banca, de seguro y de salud privada y en deportes. Las bases de datos que se usan son: temporales, documentales y relacionales.

Para solucionar problemas se usan tres pasos que son: 1 agrupar los patrones secuenciales, 2clasificarlos con los datos secuenciales y finalmente aplicar las reglas de asociación con los datos secuenciales.

Los métodos representativos son: GSP, SPADE, AprioriAll, FreeSpan, SPAM, PrefixSpan, ISM, IncSp, ISE, IncSpan.

8. 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. Se estima un modelo usando los datos recolectados y a partir de estos hacer predicciones futuras.

Existen diferentes técnicas de clasificación que son: Clasificación por inducción de árbol de decisión, clasificación bayesiana, redes neuronales, Support Vector Machines (SVM) y la clasificación que se basa en las asociaciones.

La regla de Bayes es mediante una hipótesis H sustentada para una evidencia E -> p (H|E)= (p(E|H)* p(H))/p(E) en donde p(E) representa la probabilidad del suceso y P(E|H) es la probabilidad del suceso A condicionada al suceso B. En las redes neuronales se trabaja directamente con números y si se trabaja con datos nominales se deben enumerar, estas redes consisten en tres capas: entrada, oculta y de salida. El árbol de decisión es una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Útiles para problemas que mezclen datos categóricos y numéricos.