Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Minería de Datos

Resumen de Técnicas

7° Semestre

Licenciatura en Actuaría

Nombres	Matrícula
Magdaly Rodríguez Ortiz	1815330

Profesora: Mayra Berrones

Grupo: 002

Ciudad San Nicolás de los Garza – 02 de octubre del 2020.

Técnicas de minería de datos

Como bien sabemos las tareas en minería de datos se dividen en dos categorías que son descriptivas y predictivas, cada una de ellas con distintos objetivos.

Clustering.

Esta es una técnica consiste en agrupar puntos de datos y así crear particiones basándonos en similitudes. Esta técnica tiene distintos usos como lo son: investigación de mercado, identificar comunidades, prevención de crimen, procesamiento de imágenes, entre otros.

Los tipos básicos de análisis en esta técnica son:

- Centroid Based Clustering. Cada cluster es representado por un centroide; sus cluster se construyen basados en la distancia de punto de los datos hasta el centroide, después se realizan iteraciones hasta llegar al mejor resultado. 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. En este, un cluster contiene otros clusters y en este tipo se utiliza el algoritmo Hierarchical clustering.
- Distribution Based Clustering. Cada cluster pertenece a una distribución normal, los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución y un algoritmo utilizado en este tipo es Gaussian mixture models.
- Density Based Clustering. Clusters definidos por áreas de concentración, trata de conectar puntos con una pequeña distancia y un cluster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como irregular a las áreas esparcidas entre clusters.

Existen dos tipos de métodos para la realización de dicha técnica:

- Método de k-medias. Este método es un algoritmo de clustering basado en centroides, donde k representa el número de clusters y es definido por el usuario; después de esto elegimos k datos aleatorios que pasarán a ser los centroides representativos de cada cluster, luego analizamos la distancia de cada dato al

centroide más cercano, perteneciendo a su cluster, procedemos a obtener media de cada cluster y este será el nuevo centro y finalmente repetimos el proceso hasta que los clusters no cambien.

- Método del codo. Este método consiste en graficar la reducción de la varianza total a medida que k aumenta. En un punto la reducción de la varianza no disminuirá de forma significativa entre un valor k y otro. Este punto es llamado elbow plot o codo y representa el número de k a utilizar

Reglas de asociación.

Una regla de asociación se define como una implicación del tipo: " Si A => B ", es decir, "Si A entonces B" donde A es el antecedente y B el consecuente.

Estas reglas de asociación nos permiten encontrar las combinaciones de artículos o ítems que ocurren con mayor frecuencia en una base de datos transaccional y medir la fuerza e importancia de estas combinaciones.

Estas reglas de asociación tienen un sinfín de aplicaciones y algunas de las más importantes son el definir patrones de navegación dentro de la tienda, promociones de pares de productos: hamburguesas y cátsup, soporte para la toma de decisiones, análisis de información de ventas, distribución de mercancías en tiendas, segmentación de clientes con base en patrones de compra, etc.

Existen dos diferentes tipos de reglas de asociación:

- Asociación Cuantitativa.

  Con base en los tipos de valores que manejan las reglas:

  Asociación Booleana. Asociaciones entre la presencia o ausencia de un ítem.

  Asociación Cuantitativa. Describe asociaciones entre ítems cuantitativos o atributos.

- Asociación Multidimensional.

  Con base en las dimensiones de datos que involucra una regla:

  Asociación Unidimensional. Si los ítems o atributos de la regla se referencian en una sola dimensión

  Asociación Multidimensional. Si los ítems o atributos de la regla se referencian en dos o más dimensiones.

- Asociación Multinivel.

  Con base en los niveles de abstracción que involucra la regla:

  Asociación de un nivel. Los ítems son referenciados en un único nivel de abstracción.

  Asociación Multinivel. Los ítems son referenciados a varios niveles de abstracción.

Y seguimos con las métricas utilizadas en esta técnica que son el soporte, la confianza y lift.

Soporte: el número de veces o frecuencia (relativa) con que A y B aparecen juntos en una base de datos de transacciones. Y se calcula como $\frac{A \cap B}{Total\ de\ transacciones}$

Confianza: cociente del soporte de la regla y el soporte del antecedente solamente. Y se calcula como $\frac{Soporte(\text{A=>B})}{Soporte(\text{A})}$. La confianza mide la fortaleza de la regla.

Lift: refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente y se calcula como $\frac{Soporte(\text{A=>B})}{Soporte(\text{B}) * \text{Soporte(B)}}$.

Outliers.

La detección de outliers podría ser tomada como la minería de datos anómalos, esta técnica nos ayuda al detectar un problema en datos raros o atípicos en los datos, es decir, cuando se estudia el comportamiento de los valores extremos que difieren del patrón general de la muestra. Un dato atípico es una "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".

En esta técnica nos enfrentamos a un problema que es que los datos pudiesen no ser representativos de la población, pudiendo distorsionar el comportamiento de los contrastes estadísticos; aunque también pueden ser indicadores de un segmento de los datos o la población en específico.

La podemos aplicar en aseguramiento de ingresos en las telecomunicaciones, detección de fraudes financieros, seguridad y la detección de fallas, entre otros.

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

Visualización.

Es la representación gráfica de información y datos. Se utilizan 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. Esto nos es de gran utilidad al momento de tener bases de datos extensas ya que de ese modo su entendimiento es mejor visualmente.

Existen tres tipos de visualizaciones y son:

- Elementos básicos de representación de datos.

  Algunos tipos de visualizaciones básicas son:

  Gráficas: barras, líneas, columnas, puntos, "tree maps", tarta, semi-tarta, entre otros.

  Mapas: burbujas, coropletas (o mapa temático), mapa de calor, de agregación (o análisis de drilldown).

  Tablas: con anidación, dinámicas, de drilldown, de transiciones, entre otros.

- Cuadros de mando.

  Composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas. Estos son muy utilizados en organizaciones para análisis de conjuntos de variables y toma de decisiones.

- Infografías.

  No están destinadas al análisis de variables sino a la construcción de narrativas a partir de los datos, es decir, las infografías se utilizan para contar "historias".

Es importante mencionar que la visualización de los datos es de gran importancia en todas las áreas curriculares, ya que los conjuntos de habilidades están cambiando para adaptarse a un mundo basado en los datos y es mucho más entendible cualquier información al ser analizada por el ojo humano.

Regresión.

Esta 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. Esta técnica nos ayuda a analizar la relación entre una variable dependiente ´y´ y una o varias variables independientes ´x´, encontrando así una relación matemática.

A partir de esta técnica existen dos tipos de regresión:

- Regresión lineal simple.

  En este tipo el análisis de regresión sólo se trata de una variable regresora, se llama regresión lineal simple; tiene como modelo:

$$y = \beta_o + \beta_1 x + e$$

  Estimación por mínimos cuadrados:

  La estimación de $y = \beta_o + \beta_1 x$ debe ser una recta que proporcione un buen ajuste a los datos observados. El modelo ajustado por mínimos cuadrados utiliza:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} x$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

- Regresión lineal múltiple.

  Se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos ($\beta_o, \beta_1, \dots, \beta_k$).

  Se puede relacionar la respuesta “y” con los k regresores, o variables predictivas bajo el modelo:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e$$

Algunas de sus aplicaciones son la medicina, informática, estadística, comportamiento humano, la industria, entre otros.

Clasificación.

Técnica de minería de datos más comúnmente aplicada, que organiza o mapea un conjunto de atributos por clase dependiendo de sus características.

Existen distintas técnicas de clasificación:

- Clasificación por inducción de árbol de decisión: Son una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Útiles para problemas que mezclen datos categóricos y numéricos; útiles en clasificación, agrupamiento, regresión.
- Clasificación Bayesiana: Si tenemos una hipótesis H sustentada para una evidencia E -> p(H|E)= (p(E|H)* p(H))/p(E) Donde p(A) representa la probabilidad del suceso y p(A|B) la probabilidad del suceso A condicionada al suceso B.
- Redes neuronales: Trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse.
- Support Vector Machines (SVM).
- Clasificación basada en asociaciones.

Patrones secuenciales

Se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias; 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. Es importante mencionar que hay diferentes maneras de obtener los patrones secuenciales.

Se trata de buscar 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”.

Esta técnica tiene aplicaciones en distintas áreas en medicina, biología, bioingeniería, web, análisis de mercado, distribución y comercio, aplicaciones financieras y banca, aplicaciones de seguro y salud privada, deportes, entre otras.

Para esta técnica se utilizan distintos tipos de base de datos:

Base de datos temporales.

Base de datos documentales.

Base de datos relacionales.

Resolución de problemas:

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

Predicción.

Un modelo predictivo que nos es de gran utilidad en esta técnica es el árbol de decisión, este consiste en dividir el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Para dividir el espacio muestral es necesario aplicar una serie de reglas. Al momento de dividir si una subregión contiene datos de diferentes clases, se subdivide en regiones más pequeñas.

Los árboles se pueden clasificar en dos tipos:

- Árboles de regresión. La variable respuesta ´y´ es cuantitativa.
- Árboles de clasificación. La variable respuesta ´y´ es cualitativa.

Los árboles de decisión tienen una estructura básica a seguir y consiste en que están formados por nodos y se lee de arriba hacia abajo.

Existen distintos tipos de nodos que son:

Primer nodo. Se al inicio en función de la variable más importante.

Nodos internos. Después de la primera división encontramos estos nodos, que vuelven a dividir el conjunto de datos en función de las variables.

Nodos terminales. Se encuentran en la parte inferior del esquema, función es indicar la clasificación definitiva.

Los árboles de clasificación consisten en hacer preguntas del tipo $¿xk \leq c?$ para las covariables cuantitativas o preguntas del tipo $¿xk = nivelj?$ para las covariables cualitativas.

En este tipo de árbol existen dos tipos de nodo que son:

Nodos de decisión. Tienen una condición al principio y más nodos debajo de ellos

Nodos de predicción. No tienen ninguna condición ni nodos debajo de ellos.

Cada uno de los nodos nos presenta información de tipo:

Condición: Si es un nodo donde se toma alguna decisión.

Gini: Es una medida de impureza.

Samples: Número de muestras que satisfacen condiciones para llegar a este nodo.

Value: Cuántas muestras de cada clase llegan a este nodo.

Class: Qué clase se les asigna a las muestras que llegan a este nodo.

Un 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 hiper-rectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado $\hat{y}$.