UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICO

Asignatura:

MINERÍA DE DATOS

Tema:

RESÚMEN: TÉCNICAS DE MINERIA DE DATOS

Datos.

Profesora: Mayra Cristina Berrones Reyes
Alumno: Saúl Arath Hernández Hernández
Matrícula: 1815642

Fecha: 29/09/2020

Las tareas de minería de datos se dividen generalmente en dos categorías: Descriptivas y Predictivas. El objetivo de las descriptivas es encontrar patrones que den un resumen de las relaciones ocultas dentro de los datos; descubre las características más importantes de la base de datos. Para la categoría predictiva, predice el valor de un atributo en particular basándome en los datos recolectados de otros atributos.

Dentro de las tareas **descriptivas** tenemos a:

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 uso del clustering lo podemos ver para investigaciones de mercado, identificar comunidades, prevención del crimen, procesamiento de imágenes, entre otras.
Para su transformación de los datos, utiliza variables cuantitativas, binarias y categóricas.

Los tipos de análisis básicos dentro de clustering son:

- Centroid Based Clustering: Cada cluster es representado por un centroide. Los clusters se construyen basados en la distancia de punto de los datos hasta el centroide. Se realizan varias 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 o cercanos (los puntos más cercanos están más relacionados que otros puntos más lejanos). La característica principal es que un cluster contiene a otros clusters (representan una jerarquía). Un algoritmo usado de este tipo es *Hierarchical clustering.*

- Distribution Based Clustering: En este método cada cluster pertenece a una distribución normal, La idea es que los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal. Un algoritmo de clustering perteneciente a este tipo es *Gaussian mixture models*.

- Density Based Clustering: Los clusters son definidos por áreas de concentración. Se trata de conectar puntos cuya distancia entre sí es considerada pequeña. Un cluster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como irregular a las áreas esparcidas entre clusters.

Reglas de asociación: se derivan de un tipo de análisis que extrae información por coincidencias, con el objetivo de encontrar relaciones dentro un conjunto de transacciones, en concreto, ítems o atributos que tienden a ocurrir de forma conjunta. Las 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. Tiene aplicaciones en definir patrones de navegación dentro de tiendas, soporte para la toma de decisiones, análisis de información de ventas, segmentación de clientes con base en patrones de compra, entre otras.

Los tipos de reglas de asociación son:

- 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.

Usamos métricas de interés para el proceso que son:

Soporte: Dada una regla “Si A => B”, el soporte de esta regla se define como el número de veces o la frecuencia (relativa) con que A y B aparecen juntos en una base de datos de transacciones.

Confianza: Dada una regla “Si A => B”, la confianza de esta regla es el cociente del soporte de la regla y el soporte del antecedente solamente. 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.

Detección de Outliers: Problema de la detección de datos raros o comportamientos inusuales en los datos, son denominados casualmente como “Datos atípicos”.

Aplicaciones:

- Aseguramiento de ingresos en las telecomunicaciones.
- Detección de fraudes financieros.
- Seguridad y la detección de fallas.

Método de resolución:

Se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos

Visualización de datos: es 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.

Existen multitud de técnicas y aproximaciones para la visualización según sea la naturaleza del dato de la información. Según la complejidad y elaboración de la información podemos tener la siguiente clasificación:

- Elementos básicos de representación de datos:
    - 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)
    - Tablas: con anidación, dinámicas, de drilldown, de transiciones, etc.

- Cuadros de mando: Un cuadro de mando es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas. Son ampliamente utilizados en las organizaciones para análisis de conjuntos de variables y toma de decisiones.
- Infografías: Esta narrativa no se construye a través de texto, sino mediante la disposición de la información en la que las visualizaciones se combinan con otros elementos como: símbolos, leyendas, dibujos, imágenes sintéticas, etc.

Dentro de las tareas **predictivas** tenemos a:

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

Tiene sus aplicaciones dentro de la medicina, informática, estadística, en el comportamiento humano y dentro de la industria.

Dentro de la regresión, se encuentra el modelo lineal que puede ser simple o múltiple

Cuando el análisis de regresión sólo se trata de una variable regresora, se llama regresión lineal simple. La regresión lineal simple tiene como modelo: $y = \beta_o + \beta_1 x + e$.
Un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos. $\beta_o, \beta_1, \ldots, \beta_k$ En general, se puede relacionar la respuesta "y" con los k regresores, o variables predictivas bajo el modelo: $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$.

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 entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras.

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. Problemas con la inducción de reglas: Las reglas no necesariamente forman un árbol. Las reglas pueden no cubrir todas las posibilidades. Las reglas pueden entrar en conflicto.

- 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.
  Se usan en Clasificación, Agrupamiento, Regresión; las redes neuronales consisten

generalmente de tres capas: de entrada, oculta y de salida. Internamente pueden verse como una gráfica dirigida.

- 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. Es una clase especial de dependencia en las que el orden de acontecimientos es considerado. Son eventos que se enlazan con el paso del tiempo.

- 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".
- El objetivo de la tarea es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos.
- Utiliza reglas de asociación secuenciales. - reglas que expresan patrones de comportamiento secuencial, es decir, que se dan en instantes distintos en el tiempo.

Dentro de sus características:

El orden importa Su objetivo es encontrar patrones en secuencia.
Una secuencia es una lista ordenada de itemsets, donde cada itemset es 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 problemas utilizamos las siguientes herramientas:

- Agrupamiento de patrones secuenciales.
- Clasificación con datos secuenciales.
- Reglas de asociación con datos secuenciales.

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. Tiene una relación con otras técnicas ya que 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.

Tiene sus aplicaciones en:

- Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro.
- Predecir el precio de venta de una propiedad.
- Entre otras…

Cuenta con técnicas de predicción que son las siguientes:

- Regresión lineal.
- Regresión lineal multivariante.
- Regresión no lineal.
- Regresión no lineal multivariante.
- Redes neuronales: utiliza los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión.