UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO

MATEMÁTICAS

Minería de Datos

“Resumen Técnicas Minería de Datos”

Profesor: Mayra Cristina Berrones Reyes

Alumna: Valeria Solís Agundis

Matricula: 1815413

Carrera: LA

Semestre: 7mo

Grupo:002

01/10/20

La minería de datos como ya vimos es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de los datos, es decir, es el descubrimiento eficiente de información valiosa (no obvia) de una gran colección de datos.

Dentro de esta, existen las tareas de la minería de datos que generalmente se dividen en dos, Predictivas y Descriptivas donde:

Predictivas: Predecir el valor de un atributo en particular basándome en los datos recolectados de otros atributos.

- Regresión
- Clasificación
- Patrones Secuenciales
- Predicción

Descriptivas: Su objetivo es encontrar patrones que den un resumen de las relaciones ocultas dentro de los datos. Descubre las características mas importantes de la base de datos

- Clustering
- Reglas de Asociación
- Detección de outliers
- Visualizacion

A continuación, se describirá más a detalle cada técnica mencionada.

Descriptivas

Reglas de asociación

Las 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.

Una regla de asociación se define como una implicación del tipo: " Si A => B " (antecedente y consecuencia) donde A y B son ítems individualmente. Por ejemplo: • Cereal => Leche.

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

Tipos de Reglas de Asociación

- Asociación Cuantitativa: Con base en los tipos de valores que manejan las reglas: Asociación Booleana y Cuantitativa.
- Asociación Multidimensional: Con base en las dimensiones de datos que involucra una regla: Asociación Unidimensional y Asociación Multidimensional

- Asociación Multinivel: Con base en los niveles de abstracción que involucra la regla: Asociación de un nivel y Asociación Multinivel

Métricas de Interés

- 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. Regla con bajo soporte; puede haber aparecido por casualidad.
- 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
  Regla con baja confianza; es probable que no exista relación entre antecedente y consecuente
- Lift: Refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente (Se esperaría que el valor de lift sea mayor a 1)

Visualización:

La 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. La visualización esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

Tipos de visualización: 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 (El más sencillo de todos): Gráficas, tablas, mapas.
- Cuadros de mando: Un cuadro de mando es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas.
- 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".

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.

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 puede ser empleado, por ejemplo, para caracterizar clientes, formar taxonomías, clasificar documentos, investigación de mercado, identificar comunidades, prevención de crímenes, procesamiento de imágenes, etc.

Existen diversos tipos de análisis como: Centroid based clustering, connectivity based clustering, distribution based clustering, density based clustering.

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 y 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)y un algoritmo usado de este tipo es el 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.

El método k medias que vemos que es muy implementado en este tipo de técnica se basa en centroides donde K representa el número de clusters y es definido por el usuario.

Una vez que se tiene el valor de K se procede a

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.

Detección de Outliers

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.

Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos, sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar. Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento valido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra.

- Los casos atípicos pueden clasificarse en 4 categorías:
- Casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.
- Observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento valido de la población y puede ser eliminado del análisis.
- Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son ´únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis, pero estudiando que influencia ejercen en los procesos de estimación de los modelos considerados.
- Datos extraordinarios para las que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con 1 y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el porqué de dichas observaciones.

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

Y potencialmente se pueden aplicar en el aseguramiento de ingresos en las telecomunicaciones en la detección de fraudes financieros, seguridad y la detección de fallas, entre otras ramas.

Predictivas

Regresión

La regresión es una herramienta de gran uso en el ámbito de la ingeniería y sobre todo, la estadística, el big data y data mining.

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

La regresión lineal se divide en: Regresión lineal simple y lineal múltiple.

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

La regresión tiene muchas utilidades en la vida y aplicación, siendo una de las técnicas mas aplicadas, algunos ejemplos de estas aplicaciones serían en el área de la medicina, informática, estadística, en empresas para hacer previsiones de ventas, ingresos, o evolución de mercados.

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.

Esta se encarga de:

- Emparejar o asociar datos a grupos predefinidos (aprendizaje supervisado)
- Encuentra modelos (funciones) que describen y distinguen clases o conceptos para futuras predicciones.

Básicamente se entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras. La clasificación se divide en tipos de clasificación, a continuación, se mencionarán los más importantes.

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.

Clasificación Bayesiana: es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales

Redes Neuronales: Trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse, las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.

La clasificación ha tomado gran potencia en los últimos años algunos ejemplos de aplicación serían calificación de crédito (credit scoring), reconocimiento de imágenes y patrones, diagnóstico médico, detección de fallos en aplicaciones industriales, clasificar tendencias de mercados financieros, entre otros.

Patrones Secuenciales

Los 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.

El patrón secuencial 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.

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" como, por ejemplo: si se compra una casa, 65% de las veces se comprará un refrigerador dentro de las siguientes dos semanas.

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.

Los patrones secuenciales utilizan reglas de asociación secuenciales las cuales son reglas que expresan patrones de comportamiento secuencial, es decir, que se dan en instantes distintos en el tiempo.

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.

La agrupación de los patrones secuenciales es 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.

Los patrones secuenciales tienen diversas aplicaciones en las diversas áreas como en la medicina al predecir si un químico causa cáncer, o en el análisis de mercado al estudiar el comportamiento del mercado.

Predicción

Consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento.

Existen varias maneras de hacer predicciones, desde las más sencillas a las más complejas, como lo son los árboles de decisión, los cuales 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. Estos se dividen en:

- 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.
- 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 hiper-rectángulo tendrán el mismo valor estimado $\hat{y}$

Los árboles de decisión tienen como que son fáciles de entender e interpretar, requieren poca preparación de los datos y las covariables pueden ser cualitativas o cuantitativas además no exigen supuestos distribucionales.

Una forma de mejorar un modelo predictivo es usando la técnica creada por Leo Breiman que denominó Bagging (o Bootstrap Aggregating). Esta técnica consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.