Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

MINERÍA DE DATOS

Resumen: “Técnicas de Minería de Datos”

7° Semestre

Licenciatura en Actuaría

Alumno: Evelyn Lizbeth Trejo Rodríguez

Matrícula: 1811917

Profesor: Mayra Cristina Berrones Reyes

Grupo: 002

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

TÉCNICAS DE MINERÍA DE DATOS

Las tareas de minería de datos se dividen generalmente en dos categorías:

DESCRIPTIVAS

Cuyo objetivo es encontrar patrones que den un resumen de las relaciones ocultas dentro de los datos además descubre las características más importantes de la base de datos.

Entre las técnicas descriptivas se encuentran:

Clustering

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

Algunos de sus usos principales son:

- Investigación de mercado
- Prevención de crimen
- Identificar comunidades
- Procesamiento de imágenes

Estos datos se pueden transformar en variables cuantitativas, binarias y categóricas.

<u>Hay 4 tipos de básicos de clustering</u>

- Centroid Based Clustering:

  Definimos que 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) su característica

principal es que un cluster contiene a otros clusters (representan una jerarquía) y Hierarchical clustering es el algoritmo usado para este tipo.

- 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 normal y Gaussian Mixture Models es el algoritmo de clustering perteneciente a este tipo.

- Density Based Clustering:

Los clusters son definidos por áreas de concentración y tratan 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 algoritmo de clustering está basado en centroides en donde K representa el número de clusters y es definido por el usuario. Una vez que escogemos el valor de k:

Pasos para el metodo 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:** Obtenemos la media de cada cluster y este será el nuevo centro.
4. **Iterar:** Repetimos el proceso hasta que los clusters no cambien.

Después el método del codo consiste en graficar la reducción de la varianza vs k y se toma como k al punto en que varianza no disminuirá de forma significativa entre un valor k y otro, este método se usa para saber cuál es nuestro número k de clústers óptimo.

Reglas de Asociación

Las reglas de asociación se derivan de un 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, ademas nos permiten encontrar

las combinaciones de artículos o ítems que ocurren con mayor frecuencia en una base de datos transaccional y mide la fuerza e importancia de estas combinaciones.

Una regla de asociación se define como una implicación del tipo:

" Si A => B "

Antecedente Consecuente

donde A y B son ítems individuales.

Unas de sus aplicaciones son definir patrones de navegación dentro de la tienda, promociones de pares de productos, soporte para la toma de decisions, análisis de información de ventas, distribución de mercancías en tiendas entre otros.

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.

Métricas de Interés

Las métricas de interés son utilizadas para solucionar el problema de que habían demasiadas reglas de asociación innecesarias, por lo tanto usamos:

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.

• En lenguaje de probabilidad, el soporte es:

Soporte (A=>B) = P(A∩B)

$$Frecuencia\ en\ que\ A \cap B\ aparecen\ en\ las\ transacciones\ /\ Total\ de\ transacciones$$

• El primer requisito que podemos imponer para limitar el número de reglas es que tengan un soporte mínimo.

Una regla con bajo soporte significa que 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.

Confianza (A => B) = $Soporte\ (A \Rightarrow B)$ / $Soporte\ (A)$

La confianza mide la fortaleza de la regla.

• En lenguaje de probabilidad, la confianza es una probabilidad condicional:

Confianza (A=>B) =P(B/A) = $P(A \cap B)$ / $P(A)$

Una regla con baja confianza significa que es probable que no exista relación entre antecedente y consecuente.

Lift

El Lift refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente

$Lift\ A \Rightarrow B$ = $Soporte$ (A=>B) / $Soporte\ (A) * Soporte(B)$ = $P(A \cap B)$ / $P\ (A) * P(B)$

Un lift:

>1	Representa relación fuerte y frecuencia mayor que el azar (complementos).
≈1	Representa relación del azar.
<1	Representa relación débil y frecuencia menor que el azar(sustitutos).

Detección de Outliers

Se denominan también como datos atípicos, 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

Tipos de outliers

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 válido 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 qué 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.

Aplicaciones

• Aseguramiento de ingresos en las telecomunicaciones.

• Detección de fraudes financieros.

• Seguridad y la detección de fallas.

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

Visualización

Esta técnica 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, es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

Tipos de visualizaciones

1. **Elementos básicos de representación de datos.**

   Es el caso más sencillo, a continuación, se señalan algunos tipos de visualizaciones básicas: • 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.

2. **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.

3. **Infografías**

   Las 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”. 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.

Importancia de la visualización de datos en cualquier empleo

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.

PREDICTIVAS

Son técnicas que nos ayudan a predecir el valor de un atributo en particular basándose en los datos recolectados de otros atributos.

Entre las técnicas predictivas se encuentran:

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. La regresión se encarga de analizar el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática.

Existen dos tipos de regresión, la regresión lineal simple y la múltiple.

Regresión lineal simple

Cuando el análisis de regresión sólo se trata de una variable regresora, se llama regresión lineal simple. La regresión lineal simple tiene como modelo:

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

En donde la cantidad 'e' en la ecuación es una variable aleatoria normalmente distribuida con E(e)=0 y Var(e)= σ 2

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

Un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos.

$$\beta_o, \beta_1, \dots, \beta_k$$

En general, se puede relacionar la respuesta "y" con los k regresores, o variables predictivas bajo el modelo:

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

De igual forma se estima por mínimos cuadrados.

Aplicaciones:

- Medicina
- Estadística
- Informática
- Industria
- Comportamiento humano

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 utiliza para estimar un modelo usando los datos recolectados para hacer predicciones futuras.

Técnicas de Clasificación

- **Clasificación por inducción de árbol de decisión**

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, son útiles en Clasificación, Agrupamiento y Regresión.

• **Clasificación Bayesiana**

Regla de Bayes

Si tenemos una hipótesis H sustentada para una evidencia

$$E \rightarrow p(H|E) = (p(E|H)*p(H)) / p(E)$$

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**

Algunos problemas con la inducción de reglas:

a) Las reglas no necesariamente forman un árbol.
b) Las reglas pueden no cubrir todas las posibilidades.
c) Las reglas pueden entrar en conflicto.

Patrones Secuenciales

Los patrones secuenciales describen 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 decir, se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias y es una clase especial de dependencia en las que el orden de acontecimientos es considerado.

Para 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”.

Características

• El orden importa

• Su objetivo es encontrar patrones en secuencia.

• El tamaño de una secuencia es su cantidad de elementos (itemsets).

• La longitud de una secuencia es su cantidad de ítems.

Resolución de Problemas

- Agrupamiento 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.

- Clasificación con 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.

Métodos Representativos

Existen algunos métodos representativos tales como GSP, SPAGE, AprioriAll.

Predicción

Los elementos para hacer un buen modelo de predicción son:

• Definir adecuadamente nuestro problema.

• Recopilar datos.

• Elegir una medida o indicador de éxito.

• Preparar los datos

En los modelos los datos se dividen en un 70% conjunto de entrenamiento, 15% conjunto de validación y 15% conjunto de pruebas.

Árboles Aleatorios

Árbol de decisión

Es 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.

Estructura básica de un árbol de decisión

Dentro de un árbol de decisión distinguimos diferentes tipos de nodos:

• 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 encontramos estos nodos, que vuelven a dividir el conjunto de datos en función de las variables.

• Nodos terminales u hojas: se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva.

Otro concepto es la profundidad de un árbol, que viene determinada por el número máximo de nodos de una rama.

Árbol de Clasificación

Aquí la variable respuesta es cualitativa y consiste en hacer preguntas del tipo ¿$xk \leq c$? para las covariables cuantitativas o preguntas del tipo ¿$xk = nivelj$? para las covariables cualitativas. Hay dos tipos de nodo:

• Nodos de decisión: tienen una condición al principio y tienen más nodos debajo de ellos

• Nodos de predicción: no tienen ninguna condición ni nodos debajo de ellos. También se denominan «nodos hijo»

La información de cada nodo es la siguiente:

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

➢ **Gini:** Es una medida de impureza, cuando Gini vale 0, significa que ese nodo es totalmente puro. La impureza se refiere a cómo de mezcladas están las clases en cada nodo.

Para calcular la impureza Gini, usamos la siguiente fórmula:

$$gini = 1 - \sum (probabilidad\ de\ cada\ clase)\ ^2$$

Donde pc se refiere a la probabilidad de cada clase. Podemos calcularla dividiendo el número de muestras de cada clase en cada nodo por el número de muestras totales por nodo

➢ **Samples:** Número de muestras que satisfacen las condiciones necesarias 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.

Árbol de Regresión

Este 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 y.

Algunas de las ventajas de los árboles de regresión son:

• Fácil de entender e interpretar.

• Requiere poca preparación de los datos.

• Las covariables pueden ser cualitativas o cuantitativas.

•No exige supuestos distribucionales.

Bosques Aleatorios

Técnica de aprendizaje automático supervisada basada en árboles de decision, 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.

Bagging

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.

Ventajas y Desventajas de los Bosques Aleatorios

Dado que un Bosque Aleatorio es un conjunto de árboles de decisión, y los árboles son modelos no-paramétricos, estos tienen las mismas ventajas y desventajas de los modelos no-paramétricos:

• Ventaja: pueden aprender cualquier correspondencia entre datos de entrada y resultado a predecir

• Desventaja: no son buenos extrapolando, porque no siguen un modelo conocido

Una vez terminado el modelo se debe medir su eficacia para lo que podemos usar validación cruzada la cual se emplea para estimar la tasa de error de un modelo y así evaluar su capacidad predictiva. También existen métricas de eficacia tanto para datos numéricos como categóricos como lo son:

- **Error cuadrático medio:** que mide el promedio de los errores al cuadrado.

- **Curva ROC:** nos sirve para conocer el rendimiento global de la prueba donde el eje X son los falsos positivos y el eje Y son los verdaderos positivos.