Gabriel Adrian Contreras Garcia 1752950 gpo.2 Minería de Datos.

TECNICAS DE MINERIA DE DATOS

Método de 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.

Técnicas de clasificación.

Hablaremos de algunas de las siguientes técnicas de clasificación:

-Regla de Bayes

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.

-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.

Patrones secuenciales.

La minería de datos puede definirse como 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

Las tareas de la minería de datos se dividen generalmente en dos categorías: Descriptivas y predictivas.

-Predictivas

Predicen el valor de un producto en particular basándose en los datos recolectados de otros atributos.

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.

-Son eventos que se enlazan con el paso del tiempo.

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”. 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.

Características.

-El orden importa

-Su objetivo es encontrar patrones en secuencia.

-Una secuencia es una lista ordenada de itemsets, donde cada itemsetes 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.

Resolución de problemas.

-Agrupación 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.

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 *consecuencia*

donde A y B son ítems individuales.

Algunos ejemplos:

•Cereal => Leche

•Harina => Huevo

Las reglas de asociación nos permiten:

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

Aplicaciones.

•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.

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.

compra (X, «computador») ⇒compra (X, «software contable»)

• **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.

compra (X, «zapatos») ⇒ compra (X, «calcetines»)

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.

-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.

-Lift

Refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente.

Visualización de datos

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.

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.

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 drill down).

• Tablas: con anidación, dinámicas, de drill down 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.

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.

OUTLIERS

-Datos atípicos.

Problema de la detección de datos raros o comportamientos inusuales en los datos.

Datos atípicos: “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”

-Dónde se puede aplicar?

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

- ¿Cómo se identifican los valores atípicos?

Para cada grupo de registros, o para un conjunto completo de registros, se utiliza la desviación estándar de un campo numérico específico o un múltiplo de la desviación estándar para establecer los límites superior e inferior de los valores atípicos.

Todos los registros con un valor en el campo numérico que sea superior al límite superior, o inferior al límite inferior, se consideran valores atípicos y se incluyen en los resultados de la salida.

La desviación estándar es una medida de la dispersión de un conjunto de datos; es decir, cuán dispersos están los valores. El cálculo de valores atípicos utiliza la desviación estándar de la población.

Causas comunes de los valores atípicos

Entre las causas comunes de los valores atípicos están las siguientes:

Causa	Acciones posibles
Error de entrada de datos	Corregir el error y volver a analizar los datos.
Problema del proceso	Investigar el proceso para determinar la causa del valor atípico.
Factor faltante	Determinar si no se consideró un factor que afecta el proceso.
Probabilidad aleatoria	Investigar el proceso y el valor atípico para determinar si este se produjo en virtud de las probabilidades; realice el análisis con y sin el valor atípico para ver su impacto en los resultados.

Predicción

-Metodología de la partición de datos.

Elementos para hacer un buen modelo de predicción.

- Definir adecuadamente nuestro problema (objetivo, salidas deseadas…).
- Recopilar datos.
- Elegir una medida o indicador de éxito.
- Preparar los datos (tratar con campos vacíos, con valores categóricos...)

-Arboles aleatorios.

Un 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.

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

Los árboles de decisión están formados por nodos y su lectura se realiza de arriba hacia abajo.

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.

-Bosques aleatorios

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 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.

- ¿Cómo funciona el algoritmo?

En forma resumida sigue este proceso:
-Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes sets de datos.
-Crea un árbol de decisión con cada set de datos, obteniendo diferentes árboles, ya que cada set contiene diferentes individuos y diferentes variables en cada nodo.
-Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (es decir, sin podar).
-Predice los nuevos datos usando el "voto mayoritario", donde clasificará como "positivo" si la mayoría de los árboles predicen la observación como positiva.

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.

-Usos del clustering.

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

-Tipos básicos de análisis.

-Centroid Based Clustering
Cada clúster es representado por un centroide. Los clústers 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 clústers 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 clúster contiene a otros clústers representan una jerarquía. Un algoritmo usado de este tipo es Hierarchical clustering.

-Distribution Based Clustering
En este método cada clúster 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 clústers son definidos por áreas de concentración. Se trata de conectar puntos cuya distancia entre sí es considerada pequeña. Un clúster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como irregular a las áreas esparcidas entre clústers.

-MÉTODO K-MEDIAS

Algoritmo de clustering basado en centroides. K representa el número de clústers y es definido por el usuario.

Pasos para este método:

1. Centroides: Elegimos k datos aleatorios que pasarán a ser los centroides representativos de cada clúster.

2. Distancias: Analizamos la distancia de cada dato al centroide más cercano, perteneciendo a su clúster.

3. Media: Obtener media de cada clúster y este será el nuevo centro.

4. Iterar: Repetimos el proceso hasta que los clústers no cambien.

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.

-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:

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

La cantidad 'e' en la ecuación es una variable aleatoria normalmente distribuida con $E(e)=0$ y $Var(e)= \sigma^2$

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.

$\beta_o, \beta_1, \ldots, \beta_k$

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

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

-Aplicaciones.

Las regresiones se pueden utilizar para estimar valores futuros, en algunas de las ramas donde este modelo se utiliza son: en la medicina, la informática, la estadística, la industria y el comportamiento humano.