Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

FCFM

Semestre: 7

Resumen Presentaciones

Materia:

Minería de Datos Gpo 002

Maestro:

Mayra Cristina Berrones Reyes

Alumno:

Jesús Ramón Castro Hernández

1887860

Outliers

La minería de datos anómalos es el problema de la detección de datos raros u observaciones con comportamientos inusuales con respecto al resto de los datos. Estas observaciones se desvían mucho del resto, apareciendo como un dato sospechoso que pudo ser generado por mecanismos diferentes al resto de los datos.

Aplicaciones principales:

-Aseguramiento de ingresos en las telecomunicaciones

-Detección de fraudes financieros

-Seguridad y detección de fallas

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

Se refiere a los datos atípicos. Esto se puede ver cuando cierta serie de datos sigue una tendencia, pero se tienen observaciones que se desvían del resto. Se puede aplicar en detección de fraudes, fallas o errores. Se deben realizar pruebas paramétricas para la detección de estos datos.

Regresión

La primera regresión lineal documentada fue el método de los mínimos cuadrados publicado por LaGrange. Gauss desarrolló de manera más profunda el método e incluía una versión del teorema de Gauss-Markóv. Los modelos lineales son una explicación ágil y simplificada de la realidad por parte de la matemática y estadística.

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. En una regresión se analiza el vínculo entre una variable dependiente y una o varias independientes, con una relación matemática.

Las regresiones se dividen en:

-Regresión Lineal Simple: Es cuando el análisis de regresión sólo se trata de una variable regresora, se llama regresión lineal simple. La regresión lineal simple tiene como modelo la ecuación $y = \beta_0 + \beta_1 x + e$ donde la cantidad 'e' en la ecuación es una variable aleatoria normalmente distribuida con $E(e) = 0$ y $Var(e) = \sigma^2$. Una forma de obtener de obtener estos valores ces con la estimación de mínimos cuadrados.

-Regresión Lineal Múltiple: Un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos, que es la siguiente $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$. Para este método también se aplica el método de mínimos cuadrados con su ampliación matricial para k variables.

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:

$Si\ A \rightarrow B$ donde A y B son ítems individuales.

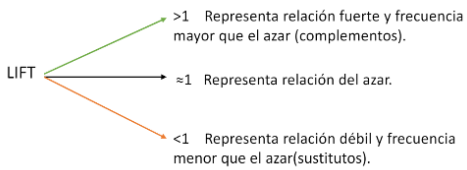
Estas reglas nos permiten:

- 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.

Las aplicaciones son como definir patrones de navegación dentro de la tienda, promociones de pares de productos, soporte para la toma de decisiones, análisis de información de ventas, distribución de mercancías en tiendas o incluso la segmentación de clientes con base en patrones de compra.

Métricas de interés:

- Soporte: 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. Una regla con soporte bajo puede haber aparecido por casualidad. Por probabilidad se define como:
  - $\frac{Frecuencia\ en\ que\ A \cap B\ aparecen\ en\ las\ atracciones}{Total\ de\ transacciones}$
- Confianza: Es el cociente del soporte de la regla y el soporte del antecedente solamente. Esta mide la fortaleza de la regla, cuando se tiene baja confianza, es probable no exista relación entre antecedente y consecuente. En probabilidad se define como:
  - $\frac{P(A \cap B)}{P(A)}$
- Lift: Refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente. Se presenta como:
  - $Lift(A \rightarrow B) = \frac{Soporte(A \rightarrow B)}{Soporte(A) * Soporte(B)} = \frac{P(A \cap B)}{P(A) * P(B)}$
  - Su valor se interpreta por los siguientes valores:



Predicción

Árbol aleatorio: 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:

Árboles de regresión en los cuales la variable respuesta y es cuantitativa.

Árboles 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 de clasificación en los cuales la variable respuesta y es cualitativa.

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.

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.

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 arboles predicen la observación como positiva.

Clustering

Es una técnica de machine learning no supervisado, es decir, la maquina aprende a través de los datos, sin embargo, no otorga una interpretación de los datos. El proceso consiste en agrupar los datos con características similares. Estos grupos se llaman clúster y son diferentes entre sí.

Usos:

- Investigación de mercado
- Identificar comunidades
- Prevención del crimen
- Procesamiento de imagen

Para usar esta técnica, las variables. Por ejemplo, cuantitativas, deben ser estandarizadas. En binarias no se requiere transformar. En variables categóricas se suele usar la binarización.

Tipos de análisis:

- Centroid Based Clustering:
    - Cada clúster se representa por un centroide. Se construyen en base a la distancia del punto al centroide. Se realizan varias iteraciones para llegar al resultado. El algoritmo más usado es K-medias.
- Connectivity Based Clustering:
    - Se basa en la premisa de la cercanía de los puntos aumenta su relación-los clústeres contienen a otros clústeres de forma jerárquica. El algoritmo mas común es Hierarchical Clustering.
- Distribución Based Clustering
    - Cada clúster pertenece a una distribución normal. Se basa en la probabilidad de cada punto de pertenecer a la misma distribución. El algoritmo común es el Gaussian Mixture Models.
- Density Based Clustering:
    - Se trata de conectar puntos cuya distancia entre si es pequeña. Se define por áreas. El clúster contiene a todos los puntos relacionados dentro de una distancia limitada.

Visualización

Es la representación gráfica de los datos y la información. Se divide en tres:

- Elementos básicos de representación de datos. (Tablas, gráficas o mapas)
- Cuadros de mando es una composición compleja de visualizaciones individuales que guardan coherencia y una relación temática entre ellas. (Se usan para analizar 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. Esta narrativa no se construye a través de texto, sino mediante la disposición de la información combinando símbolos, leyendas, dibujos, etc.
-

Estándares de visualización de datos, como HTML5, CSS3, SCV, entre otros.

La importancia f}de esto es de alto grado. Dado que, conforme va avanzando el tiempo, se tiene que analizar de forma diferente. La toma de decisiones se ve beneficiada con la correcta visualización de la información, así como facilita la interpretación.

Patrones Secuenciales

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 en un tiempo dado.

Se trata de buscar asociaciones de la forma de "si sucede el evento x en el tiempo t, sucederá y en el tiempo t+1".

- El orden importa
- Tiene como objetivo encontrar patrones
- El tamaño de la secuencia es la cantidad de ítems en la misma.

Áreas de aplicación:

- Medicina
- Biología
- Web
- Mercados
- Finanzas
- Etc.

Ayuda a resolver el agrupamiento de patrones secuenciales, que se define como la tarea de separar en grupos a los datos, de manera que un grupo sean muy similares entre sí, Al mismo tiempo sean diferentes a los otros grupos.

Reglas de asociación con datos secuenciales:

Se presentan cuando los datos contiguos presentan algún tipo de relación.

Métodos: (GSP, SPADE, APRIORIALL, FREESPAN, SPAN, ISM, ISE, INCSPAIN)

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. El proceso consiste en estimar un modelo usando los datos recolectados para hacer predicciones futuras.

Las técnicas más comunes son:

-Por árbol de decisión: Son una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Útiles para problemas que mezclen datos categóricos y números. Útiles en clasificación, agrupación y regresión.

-Clasificación Bayesiana: Si tenemos una hipótesis H sustentada para una evidencia $E \rightarrow P(H|E) = \frac{P(E|H)*P(H)}{P(E)}$ donde la P(A) representa la probabilidad del suceso y P(A|B) la probabilidad del suceso A condicionada al suceso B.

-Redes Neuronales: Las redes neuronales permiten buscar la combinación de parámetros que mejor se ajusta a un determinado problema. Las redes neuronales son un modelo para encontrar esa combinación de parámetros y aplicarla al mismo tiempo. En el lenguaje propio, encontrar la combinación que mejor se ajusta es "entrenar" la red neuronal. Una red ya entrenada se puede usar luego para hacer predicciones o clasificaciones, es decir, para "aplicar" la combinación.

-Support Vector Machines (SVM)

-Clasificación basada en asociaciones