Universidad Autónoma de Nuevo León

Licenciatura en Actuaría

Técnicas de Minería de Datos

Suarez Martinez Tadeo Alejandro
7to Semestre
1806069

02 de Octubre del 2019

Descriptivas

Clustering

El clustering es una técnica la cual es muy conocida del aprendizaje no supervisado, es decir es un modelo que se ajusta a las observaciones, consiste en agrupar puntos de datos y de esta forma crear particiones basándonos en similitudes, algunos usos del clustering podría ser en investigación de mercado, identificar comunidades, prevención del crimen y procesamiento de imágenes, si los datos son categóricos tienen que binarizarse o estandarizarse en caso de ser cuantitativos, entendimos que existen 4 tipos básicos del clustering:

1. Centroid Based Clustering.
2. Connectivity Based Clustering.
3. Distribution Based Clustering.
4. Density Based Clustering.

DISTRIBUTION BASED CLUSTERING

En este método se busca que los puntos sean divididos con base en la probabilidad de pertenecer a la misma distribución normal, el algoritmo Gaussian mixture models es muy usado en este ramo.

DENSITY BASED CLUSTERING

Se trata de conectar puntos cuya distancia entre sí es considerada pequeña.

K means

- Elegimos k datos aleatorios que pasarán a ser los centroides representativos de cada cluster.
- Analizamos la distancia de cada dato al centroide más cercano, perteneciendo a su cluster.
- Obtener media de cada cluster y este será el nuevo centro.
- Repetimos el proceso hasta que los clusters no cambien.

Si deseamos saber cual es el número de clústers óptimo consiste en graficar la reducción de la varianza total a medida que k aumenta.

Reglas de asociación

En este método descriptivo se deriva 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.
- Medir la fuerza e importancia de estas combinaciones.

Tipos de Reglas de Asociación

- Asociación Booleana: asociaciones entre la presencia o ausencia de un ítem.
- Asociación Cuantitativa: describe asociaciones entre ítems cuantitativos o atributos.
- Asociación Unidimensional: los ítems o atributos de la regla se referencian en una sola dimensión.
- Asociación Multidimensional: los ítems o atributos de la regla se referencian en dos o más dimensiones.
- 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.

Algoritmo Apriori

Para aplicar este algoritmo tendremos estos conceptos:

Soporte: número de veces o frecuencia (relativa) con que A y B aparecen juntos en una base de datos de transacciones.

$$Soporte(A \rightarrow B) = P(A \cap B) = Frecuencia\ en\ que\ A \cap B\ aparece\ /\ Total\ de\ transacciones$$

Una 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.

$$Confianza(A \rightarrow B) = Soporte(A \rightarrow B)\ /Soporte(A) = P(A|B) = P(A \cap B)\ P(A)$$

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 sabemos que ocurrió el antecedente.

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

- lift > 1 representa relación fuerte y de frecuencia mayor que el azar (complementos).
- lift = 1 representa relación del azar.
- lift < 1 representa relación débil y de frecuencia menor que el azar (sustitutos).

Paso 1

Se establece los valores mínimos para el soporte y la confianza.

Paso 2

Se toman todos los subconjuntos de transacciones que tienen todo un soporte mayor al del soporte mínimo.

Paso 3

Tomar todas las reglas de estos subconjuntos que tengan una confianza mayor al valor de la confianza mínima.

Paso 4

Ordenar las reglas de forma decreciente en base al valor del lift.

Detección de outliers

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

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.

Donde se puede aplicar la detección de estos datos anómalos:

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

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

Visualizacion

¿Qué es la visualización de datos?

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.

1. Elementos básicos de representación de datos: es el caso más sencillo.

a. Gráficas: barras, líneas, columnas, puntos, “tree maps”, tarta, semi-tarta, etc.

b. Mapas: burbujas, coropletas (o mapa temático), mapa de calor, de agregación (o análisis de drilldown).

c. Tablas: con anidación, dinámicas, de drilldown, de transiciones, etc.

2. Cuadros de mando: composición compleja de visualizaciones individuales que tienen coherencia y relación temática entre ellas. Son utilizados para análisis de conjuntos de variables y toma de decisiones.

3. Infografías: no están destinadas al análisis de variables sino a la construcción de narrativas a partir de los datos, es decir, se utilizan para contar “historias”. Esta

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

Regresión

Este es un tema que la mayoría de los actuarios ya dominamos perfectamente, sin embargo ahora encontraremos la manera de aplicarlo en la minería de datos, la regresión es un 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.

Como sabemos la regresión lineal simple tiene como modelo: $y = \beta_o + \beta_1 x + e$, la cantidad ‘e’ en la ecuación es una variable aleatoria la cual tiene una distribución de probabilidad normal como propiedades que $E(e)=0$ y $Var(e)= \sigma 2$, en general, se puede relacionar la respuesta “y” con los k regresores, o variables predictivas.

Aplicaciones

Realmente este método no tiene limitaciones en cuanto a campos donde se pueda aplicar, ya que estos patrones de tendencia lineal se pueden encontrar en cualquier lado, sin embargo, algunos son:

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

Clasificación

¿Qué es la 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.

¿En donde funciona?

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

Técnicas de clasificación

Hablaremos de algunas de las siguientes técnicas de clasificación:

• Clasificación por inducción de árbol de decisión

• Clasificación Bayesiana

• Redes neuronales

• Support Vector Machines (SVM)

• Clasificación basada en asociaciones

Regla de Bayes

El teorema de **Bayes** parte de una situación en la que es posible conocer las probabilidades de que ocurran una serie de sucesos Ai. A esta se añade un suceso B cuya ocurrencia proporciona cierta información, porque las probabilidades de ocurrencia de B son distintas según el suceso Ai que haya ocurrido.

Redes Neuronales

son un modelo computacional vagamente inspirado en el comportamiento observado en su homólogo biológico. Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales.

• 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 grafica dirigida.

Arboles de decisión

Los árboles de decisión son un tipo de algoritmo que clasifica la información de forma que, como resultado, se genere un modelo en forma de árbol. Se trata de un modelo esquematizado de la información que representa las diferentes alternativas junto con los posibles resultados para cada alternativa elegida. Los árboles de decisión son un tipo de modelo muy utilizado debido a que facilita mucho la comprensión de las diferentes opciones.

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

Este tema es parecido a el clustering ya que se trata de encontrar patrones o tendencias en secuencias de datos, por lo que ambos son parte del aprendizaje no supervisado a minería de secuencias es un caso particular de la minería de datos estructurados, el patrón secuencial describe por ejemplo 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.

- Buscamos asociaciones “si sucede el evento X en el instante de tiempo t entonces sucederá el evento Y en el instante t+n”.

Con esto buscamos patrones en secuencia, por lo tanto el orden importa, el tamaño de una secuencia es su cantidad de elementos, la longitud de una secuencia es su cantidad de ítem.

Predicción

Para un modelo de predicción tenemos que tomar en cuenta algunos elementos para que hacer un buen modelo:

- 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..)

Árbol de decisión

Los árboles de decisión son un tipo de algoritmo que clasifica la información de forma que, como resultado, se genere un modelo en forma de árbol. Se trata de un modelo esquematizado de la información que representa las diferentes alternativas junto con los posibles resultados para cada alternativa elegida. Los árboles de decisión son un tipo de modelo muy utilizado debido a que facilita mucho la comprensión de las diferentes opciones.

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.

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.

Árbol de regresión

Los árboles de decisión, donde la variable de destino puede tomar valores continuos (por lo general números reales).

Bosques aleatorios

Combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos, Su principal ventaja es que obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar.