Universidad Autónoma de
Nuevo León

Facultad de Ciencias Físico – Matemáticas

Minería de Datos

Profesor
Mayra Cristina Berrones Reyes

Tarea
Resumen de técnicas

Alumna
Melany Salazar Mata

Grupo 002 **Matricula** 1679234

02 de octubre del 2020

Técnicas de minería de datos

Descriptivas

CLUSTERING: Técnica de aprendizaje de maquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear particiones basándonos en similitudes. Se puede utilizar en: Investigación de mercado, Identificar comunidades, Prevención de crimen, Procesamiento de imágenes.

Tipos básicos de análisis:

- ✓ 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). El algoritmo más usado de este tipo es el de k-medias.
- ✓ Connectivity Based Clustering: Los clusters se definen agrupando los datos mas similares o cercanos. Un cluster contiene a otros clusters. El algoritmo usado de este tipo es Hierarchical clustering.
- ✓ Distribution Based Clustering: Cada cluster pertenece a una distribucion normal. La idea es que los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución. Un algoritmo es Gaussian mixture models.
- ✓ Density Based Clustering: Los clusters son definidos por áreas de concentración. Conectan puntos cura distancia entre si es considerada pequeña.

K representa el numero de clusters y es definido por el usuario.
La varianza de cada cluster disminuye al aumentar k. Si sólo hay un elemento en el cluster, la varianza es de 0. Entre menor sea la suma de las varianzas de los clusters, mejor es nuestro clustering.

REGLAS DE ASOCIACIÓN: Se derivan de un tipo de análisis que extrae información por coincidencias, con el objetivo de encontrar relaciones dentro de un conjunto de transacciones, en concreto; ítems o atributos que tienden a ocurrir de forma conjunta. Descubrir hechos que ocurren en común dentro de un conjunto de datos. Relación relevante entre los elementos.

Las 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.

Tipos de reglas:

- ❖ 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.

DETECCIÓN DE OUTLIERS: Estudia el comportamiento de valores extremos que difieren del patrón general de una muestra. Problema de la detección de datos raros o comportamientos inusuales en los datos.

Se puede aplicar en:

- 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.

Los Outliers pueden significar varias cosas:

1. ERROR: Si tenemos un grupo de “edades de personas” y tenemos una persona con 160 años, seguramente sea un error de carga de datos. En este caso, la detección de outliers nos ayuda a detectar errores.
2. LIMITES: En otros casos, podemos tener valores que se escapan del “grupo medio”, pero queremos mantener el dato modificado, para que no perjudique al aprendizaje del modelo de ML.
3. Punto de Interés: puede que sean los casos “anómalos” los que queremos detectar y que sean nuestro objetivo (y no nuestro enemigo!)

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.

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

Tipos de visualizaciones: 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:

1. Elementos básicos de representación de datos: Algunos tipos de visualizaciones básicas:
    - Gráficas
    - Mapas
    - Tablas
2. Cuadros de mando: Es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas
3. 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.

Predictivas

REGRESIÓN: 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 $Var(e)= \sigma^2$.

Regresión lineal múltiple:

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

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

CLASIFICACIÓN: Técnica de minería de datos más comúnmente aplicada, que organiza o mapea un conjunto de atributos por clase dependiendo de sus características.
¿Para qué puede funcionar? Se estima un modelo usando los datos recolectados para hacer predicciones futuras.
Técnicas de Clasificación:

- ✓ Clasificación por inducción de árbol de decisión
- ✓ Clasificación Bayesiana
- ✓ Redes neuronales
- ✓ Support Vector Machines (SVM)
- ✓ Clasificación basada en asociaciones

Redes Neuronales: Trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse.
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 SECUANCIALES: 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.
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 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.

Aplicaciones: Medicina, Biología, Bioingeniería, Web, Análisis de mercado, distribución y comercio, Aplicaciones financieras y banca, Aplicaciones de seguro y salud privada, Deportes.

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.

Reglas de asociación con datos secuenciales: Se presenta cuando los datos contiguos presentan algún tipo de relación.

Clasificación con datos secuenciales: Éstos expresan patrones de comportamiento secuenciales, es decir que se dan en instantes distintos (pero cercanos) en el tiempo.

PREDICCIÓN: Tenemos que definir adecuadamente nuestro problema (objetivo, salidas deseadas,etc)., recopilar datos, elegir una medida o indicador de éxito y preparar los datos.

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.

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.
- ➢ 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.

Random Forest: 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.