



UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN



FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

RESUMEN DE LAS TÉCNICAS DE MINERÍA DE DATOS

MINERIA DE DATOS

NOMBRE: ALISON LEONOR ROLDAN SANCHEZ

MATRICULA: 1806488

MAESTRO: MTS Mayra Cristina Berrones R.

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

DESCRIPTIVO:

- Clustering
- Reglas de asociación
- Detección de outliers
- Visualización

PREDICTIVO:

- Regresión
- Clasificación
- Patrones secuenciales
- Predicción

DESCRIPTIVO

CLUSTERING

DEF. 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:

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

Transformación de datos:

- ❖ Variables cuantitativas
- ❖ Variables binarias
- ❖ Variables categóricas

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. 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 clusters 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 cluster contiene a otros clusters (representan una jerarquía). Un algoritmo usado de este tipo es Hierarchical clustering

- Distribution Based Clustering

En este método cada cluster 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 clusters son definidos por áreas de concentración. Se trata 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.

REGLAS DE ASOCIACION

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 “

A: antecedente

B: consecuencia, donde A y B son ítems individuales.

Algunos ejemplos: • Cereal => Leche • Harina => Huevo

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 Cátup.
- 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.
- 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.

DETECCION DE OUTLIERS

Minería de 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

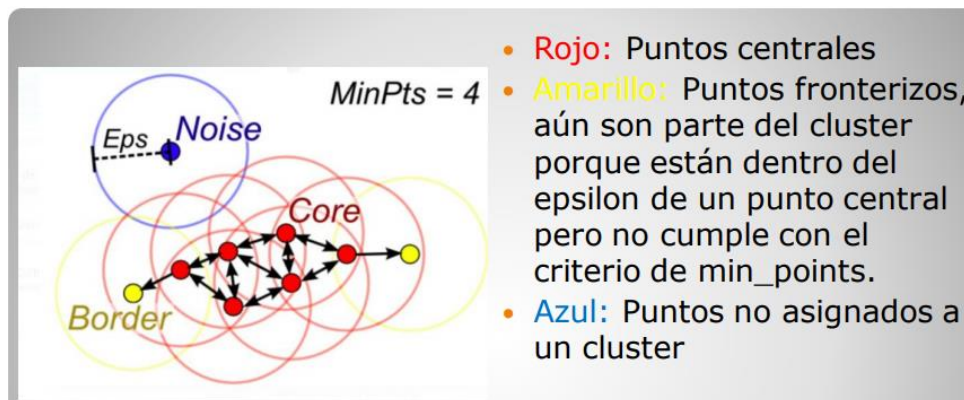
Aplicaciones:

- ✓ 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.

Ejemplo:

Teniendo como muestra los salarios de las personas en una región delimitada, nos encontraríamos con salarios muy similares, sin embargo, sabemos que existirán directivos de empresas o funcionarios públicos, con un salario excesivamente alto, comparado con los demás. Es a este dato al que conoceremos como Outlier. Este dato tan alto afectaría directamente a la media, por lo que, en estos casos, la opción más viable sería obtener la mediana del conjunto de datos.



VISUALIZACION

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

Tipos de visualizaciones:

- 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.

- 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.

PREDICTIVO

REGRESION

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.

Existe la regresión lineal simple y la regresión lineal 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:

$$y = \beta_0 + \beta_1 x + e$$

La cantidad 'e' en la ecuación es una variable aleatoria normalmente distribuida con

$$E(e)=0 \text{ y } \text{Var}(e)=\sigma^2$$

Estimación por mínimos cuadrados

La estimación de $y = \beta_0 + \beta_1 x$ debe ser una recta que proporcione un buen ajuste a los datos observados. El modelo ajustado por mínimos cuadrados utiliza:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 x$$

- 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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e$$

También se estima por mínimos cuadrados

Aplicaciones:

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

CLASIFICACION

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.

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

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.

❖ Clasificación Bayesiana

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.

- ✓ 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.

❖ Support Vector Machines (SVM)

❖ Clasificación basada en asociaciones

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.

- 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

- ❖ Web
- ❖ Análisis de mercado, distribución y comercio
- ❖ Aplicaciones financieras y banca
- ❖ Aplicaciones de seguro y salud privada

Tipos de bases de datos

☐ Base de datos temporales ☐ Base de datos documentales ☐ Base de datos relacionales

Resolución del problema

- **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.

PREDICCIÓN

Metodología de la partición de datos

Para hacer un buen modelo de predicción se necesitan los siguientes elementos:

1. Definir adecuadamente nuestro problema (objetivo, salidas deseadas.....).
2. Recopilar datos
3. Elegir una medida o indicador de éxito
4. Preparar los datos
5. Dividir los datos:
 - 70% CONJUNTO DE ENTRENAMIENTO
 - 15% CONJUNTO DE VALIDACIÓN
 - 15% CONJUNTO DE PRUEBAS.

Arboles aleatorios

ARBOL DE DECISIÓN:

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.

ARBOL DE CLASIFICACION:

Consiste en hacer preguntas del tipo $\{x_k \leq c\}$ para las covariables cuantitativas o preguntas del tipo $\{x_k = nivel_j\}$ para las covariables cualitativas, de esta forma el espacio de las covariables es dividido en hiperrectángulos y todas las observaciones que queden dentro de un hiperrectángulo tendrán el mismo valor grupo estimado.

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. A continuación, veremos cómo se calcula.
- 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.

Gini, medida de limpieza

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.

ARBOL DE REGRESION:

Consiste en hacer preguntas de tipo $\{x_k \leq c\}$ para cada una de las covariables, de esta forma el espacio de las covariables es dividido en hiperrectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado y.

Ventajas:

- ✓ 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 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.

Ventajas y desventajas:

Dato que un random forest es un conjunto de árboles de decisión, y los árboles son modelos no-paramétricos, los random forests 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

Diferencia entre un árbol de decisión y un bosque aleatorio

En la siguiente imagen puedes ver la diferencia entre el modelo aprendido por un árbol de decisión y un random forest cuando resuelven el mismo problema de regresión. Este random forest en particular, utiliza 100 árboles.