UNIVERSIDAD AUTÓNOMA "GABRIEL RENÉ MORENO"

FACULTAD DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN Y TELECOMUNICACIONES

INGENIERÍA EN SISTEMAS

PROCESO A LA TOMA DE DECISIONES

MATERIA	**:**	SISTEMAS PARA EL SOPORTE A LA TOMA DE DECISIONES
SIGLA	**:**	INF423
ESTUDIANTE	**:**	Fernando Jose Ajhuacho Cahuasiri
REGISTRO	**:**	220046387
FECHA	**:**	30/09/2024
DOCENTE	**:**	Ing. Miguel Jesus Peinado Pereira

Santa Cruz de la Sierra – Bolivia

Agrupación

La técnica de agrupación (también conocida como clustering) es un método de minería de datos que consiste en dividir un conjunto de datos en grupos (o clusters) de manera que los elementos dentro de un mismo grupo sean más similares entre sí que con los de otros grupos. La agrupación es una técnica no supervisada, lo que significa que los datos no están previamente etiquetados, y su objetivo es descubrir estructuras ocultas.

Aplicaciones comunes de la agrupación:

Segmentación de clientes: Identificar grupos de clientes con comportamientos de compra similares para mejorar estrategias de marketing.

Análisis de imágenes: Agrupar píxeles o características de imágenes para la clasificación de objetos.

Detección de anomalías: Identificar puntos de datos que no encajan en ningún grupo, lo cual puede indicar fraudes o errores.

Agrupación de documentos: Organizar textos o artículos en grupos temáticos.

Genómica: Clasificar genes con funciones similares o encontrar patrones en el ADN.

Técnicas de agrupación más utilizadas:

K-means:

Un algoritmo simple y popular que agrupa los datos en "K" clusters basados en la cercanía de los puntos a los centros (centroides). Es útil cuando se conoce de antemano cuántos grupos se buscan.

Algoritmo de agrupación jerárquica:

Este algoritmo crea una jerarquía de clusters. Puede ser aglomerativo (empezando con cada dato como un cluster y combinándolos gradualmente) o divisivo (empezando con todos los datos en un solo cluster y dividiéndolos progresivamente).

DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

Agrupa puntos basados en la densidad de puntos cercanos, lo que le permite detectar clusters de formas arbitrarias y manejar datos con ruido.

Mean-shift:

No requiere que se defina el número de clusters con antelación, sino que busca densidades locales en los datos para encontrar los centros de los clusters.

Clustering difuso (Fuzzy C-means):

A diferencia de los métodos tradicionales, donde cada dato pertenece solo a un cluster, el clustering difuso asigna a cada punto un grado de pertenencia a diferentes clusters. Es útil cuando las fronteras entre grupos no son claras.

Proceso general de la agrupación:

Selección de características: Definir las variables o atributos que describen los datos.

Selección del algoritmo: Escoger el algoritmo de agrupación más adecuado según el tipo de datos y el objetivo del análisis.

Ejecución del algoritmo: Aplicar el algoritmo a los datos para identificar los clusters.

Evaluación de la calidad: Utilizar métricas como la inercia (en K-means), el coeficiente de silueta, o el índice de Davies-Bouldin para evaluar qué tan buenos son los clusters.

Interpretación y acción: Analizar los resultados para tomar decisiones o descubrir nuevos patrones.

Ventajas:

Descubre estructuras ocultas sin necesidad de supervisión.

Flexibilidad para ser aplicada en distintos dominios y tipos de datos (imágenes, texto, transacciones).

Desventajas:

Los resultados pueden ser sensibles a la elección de parámetros (como K en K-means).

Algunos algoritmos no son adecuados para datos de formas complejas o con ruido.

La interpretación de los clusters puede ser subjetiva.

La técnica de agrupación es fundamental en el análisis exploratorio de datos, ayudando a encontrar relaciones y patrones ocultos en un conjunto de datos.

Asociación

La técnica de asociación es un método en la minería de datos que busca descubrir relaciones o dependencias entre variables en grandes conjuntos de datos. Es particularmente útil para encontrar patrones de co-ocurrencia, es decir, para identificar qué elementos tienden a aparecer juntos con frecuencia. El ejemplo más famoso es el análisis de cesta de la compra, donde se descubren productos que los clientes tienden a comprar juntos, como pan y mantequilla.

Aplicaciones comunes de la técnica de asociación:

Análisis de cesta de la compra:

Utilizado por comercios para identificar productos que los clientes suelen comprar juntos, con el fin de optimizar la disposición en tiendas o sugerir productos adicionales.

Recomendaciones de productos:

Plataformas de comercio electrónico utilizan la asociación para recomendar productos relacionados basándose en compras anteriores de los usuarios.

Detección de fraudes:

En el análisis de transacciones financieras, las reglas de asociación pueden ayudar a identificar patrones de comportamiento sospechoso.

Sistemas de diagnóstico médico:

Descubrir qué síntomas o enfermedades suelen aparecer juntos, ayudando a mejorar los diagnósticos clínicos.

Algoritmos más utilizados en la técnica de asociación:

Algoritmo Apriori:

Es uno de los algoritmos más conocidos para descubrir reglas de asociación. Funciona identificando primero conjuntos de elementos frecuentes (frequent itemsets) y luego generando reglas de asociación a partir de esos conjuntos.

Ejemplo: si se detecta que los productos "leche" y "pan" suelen comprarse juntos, la regla de asociación podría ser "si se compra leche, hay una alta probabilidad de que se compre pan".

Algoritmo Eclat (Equivalence Class Clustering and Bottom-Up Lattice Traversal):

Similar a Apriori, pero en lugar de buscar combinaciones de elementos, usa una representación basada en conjuntos de transacciones para hacer el proceso de búsqueda más eficiente en ciertos tipos de datos.

Algoritmo FP-Growth (Frequent Pattern Growth):

Este algoritmo utiliza una estructura de árbol (FP-Tree) para representar los elementos de manera eficiente y reducir el número de combinaciones a explorar. Es más rápido que Apriori en muchos casos porque evita generar combinaciones innecesarias.

Conceptos clave en la técnica de asociación:

Soporte (Support):

Indica con qué frecuencia ocurre un conjunto de elementos en los datos.

Ejemplo: si "pan" y "leche" se compran juntos en el 15% de las transacciones, el soporte de esta combinación es 0.15.

Confianza (Confidence):

Mide cuán a menudo la regla de asociación es verdadera. Es la proporción de transacciones que contienen el antecedente (por ejemplo, "leche") en las cuales también se encuentra el consecuente (por ejemplo, "pan").

Ejemplo: si en el 80% de las transacciones donde se compra leche también se compra pan, la confianza de la regla "si compras leche, también compras pan" es 0.80.

Lift:

Evalúa la fuerza de una regla de asociación. Mide cuánto más probable es que dos elementos ocurran juntos que si fueran independientes. Si el lift es mayor que 1, la asociación es significativa.

Ejemplo: un lift de 1.2 indica que la compra de "pan" junto con "leche" es un 20% más probable que si fueran independientes.

Proceso general de la técnica de asociación:

Definir el conjunto de datos: Seleccionar el conjunto de transacciones o eventos donde se quiere buscar asociaciones.

Calcular los itemsets frecuentes: Utilizar algoritmos como Apriori o FP-Growth para identificar combinaciones de elementos que ocurren juntos con frecuencia.

Generar reglas de asociación: A partir de los itemsets frecuentes, generar reglas que relacionen los elementos, como "si A ocurre, entonces B es probable que ocurra".

Evaluar las reglas: Utilizar las métricas de soporte, confianza y lift para determinar la relevancia de las reglas descubiertas.

Aplicar los resultados: Usar las reglas para mejorar estrategias comerciales, sugerir recomendaciones, detectar patrones en datos, etc.

Ejemplo práctico:

En un análisis de cesta de la compra, un conjunto de transacciones puede ser algo como:

Transacción 1: {leche, pan, mantequilla}

Transacción 2: {pan, huevos}

Transacción 3: {leche, pan}

Transacción 4: {pan, mantequilla}

Al aplicar la técnica de asociación, podríamos descubrir una regla como:

Regla de asociación: Si un cliente compra pan, hay un 70% de probabilidad de que también compre mantequilla (basado en el soporte y la confianza de los datos).

Ventajas:

Permite descubrir relaciones significativas que no son evidentes a simple vista.

Puede aplicarse a distintos tipos de datos (transacciones, clics en un sitio web, etc.).

Útil para mejorar estrategias comerciales y ofrecer recomendaciones personalizadas.

Desventajas:

Puede generar un gran número de reglas, lo que complica su análisis e interpretación.

No siempre es fácil seleccionar los parámetros adecuados (como el soporte mínimo) para obtener resultados útiles.

Las asociaciones descubiertas no implican causalidad, solo correlaciones.

La técnica de asociación es una herramienta poderosa para descubrir patrones de co-ocurrencia en los datos, lo que permite a las empresas y organizaciones tomar decisiones más informadas basadas en el comportamiento de los usuarios o clientes.

Clasificación

La técnica de clasificación es un método de minería de datos supervisada que tiene como objetivo asignar una etiqueta o clase a un conjunto de datos en función de sus características. A diferencia de otras técnicas, en la clasificación se trabaja con un conjunto de datos etiquetado, es decir, ya se conocen las clases o categorías de los datos, y el objetivo es predecir a qué clase pertenecerán nuevos datos no etiquetados.

Aplicaciones comunes de la clasificación:

Detección de fraudes:

Clasificar transacciones como legítimas o fraudulentas basándose en su comportamiento.

Diagnóstico médico:

Predecir si un paciente tiene o no una enfermedad específica en función de los síntomas y los resultados médicos.

Reconocimiento de imágenes:

Clasificar imágenes en categorías como "gato", "perro", "coche", etc.

Filtrado de correo:

Clasificar correos electrónicos como "spam" o "no spam".

Crédito y préstamos:

Predecir si un solicitante de crédito tiene un riesgo alto o bajo basándose en su historial financiero.

Algoritmos de clasificación más utilizados:

Árboles de decisión:

Un algoritmo gráfico que utiliza un modelo jerárquico de decisiones basado en reglas. Los nodos representan características, las ramas representan decisiones, y las hojas representan clases o resultados.

Ventaja: Fácil de interpretar y visualizar.

Ejemplo: Un árbol de decisión puede usar características como el ingreso y la edad para clasificar a una persona como un buen o mal solicitante de crédito.

K-Nearest Neighbors (K-NN):

Un algoritmo simple que clasifica un dato basándose en las clases de los K vecinos más cercanos en el espacio de características.

Ventaja: Fácil de implementar y no requiere suposiciones sobre la distribución de los datos.

Ejemplo: En un conjunto de datos de clasificación de frutas, si un nuevo punto de datos tiene características similares a tres manzanas y dos naranjas, el sistema lo clasificará como "manzana".

Máquinas de soporte vectorial (SVM):

Este algoritmo intenta encontrar el hiperplano que mejor separa las diferentes clases en un espacio de características. Es especialmente útil para problemas donde las clases son linealmente separables.

Ventaja: Es efectivo en espacios de alta dimensionalidad y es muy robusto para problemas complejos.

Ejemplo: Separar correos electrónicos "spam" y "no spam" con un hiperplano basado en características como la frecuencia de ciertas palabras.

Redes neuronales:

Inspiradas en el cerebro humano, las redes neuronales están formadas por capas de neuronas conectadas entre sí. Son especialmente útiles en problemas complejos como el reconocimiento de imágenes o el procesamiento del lenguaje natural.

Ventaja: Capaces de aprender patrones muy complejos y no lineales.

Ejemplo: Clasificar imágenes en categorías como "gatos" o "perros" tras entrenarse con millones de ejemplos.

Naive Bayes:

Basado en el teorema de Bayes, este algoritmo asume que todas las características son independientes entre sí. A pesar de ser una suposición fuerte, funciona sorprendentemente bien en muchos casos, como el filtrado de correo spam.

Ventaja: Simple y rápido, con buenos resultados en clasificación de texto.

Ejemplo: Clasificar correos electrónicos como "spam" o "no spam" en función de la probabilidad de ciertas palabras en el texto.

Regresión logística:

Aunque se llama "regresión", este algoritmo se utiliza para problemas de clasificación binaria. Modela la probabilidad de que un evento ocurra en función de las variables predictoras.

Ventaja: Fácil de interpretar y eficaz para problemas de clasificación binaria.

Ejemplo: Predecir si un cliente comprará un producto basado en características como edad, ingresos y comportamiento de compra anterior.

Proceso general de la técnica de clasificación:

Recolección de datos:

Se recolectan los datos que contienen las características de interés y sus respectivas etiquetas (clases).

Preprocesamiento de datos:

Se limpian los datos, se eliminan las inconsistencias y se transforman si es necesario (normalización, escalado, codificación de variables categóricas).

División de datos:

Se dividen los datos en dos subconjuntos: un conjunto de entrenamiento y uno de prueba. El conjunto de entrenamiento se usa para ajustar el modelo, mientras que el conjunto de prueba se utiliza para evaluar su rendimiento.

Entrenamiento del modelo:

Se selecciona un algoritmo de clasificación y se entrena con los datos etiquetados.

Evaluación del modelo:

Se mide el rendimiento del modelo utilizando el conjunto de prueba y métricas como precisión, exactitud (accuracy), recuperación (recall), precisión (precision) y la curva ROC.

Ajuste de hiperparámetros:

Si es necesario, se ajustan los parámetros del algoritmo para mejorar su rendimiento.

Predicción:

Una vez entrenado y evaluado, el modelo se utiliza para clasificar nuevos datos no etiquetados.

Métricas comunes de evaluación:

Precisión (Accuracy):

Es el porcentaje de predicciones correctas en relación con el total de predicciones. Es una métrica general, pero puede ser engañosa si las clases están desbalanceadas.

Precisión (Precision):

Mide la proporción de predicciones positivas correctas frente al total de predicciones positivas. Es útil cuando los falsos positivos son costosos.

Recuperación (Recall):

Mide la proporción de verdaderos positivos que fueron correctamente identificados frente al total de verdaderos positivos. Es útil cuando los falsos negativos son costosos.

F1-score:

Es la media armónica entre la precisión y la recuperación, y proporciona un equilibrio entre ambas métricas cuando el costo de falsos positivos y negativos es similar.

Curva ROC y AUC:

La curva ROC muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos a varios umbrales de decisión. El Área Bajo la Curva (AUC) mide el rendimiento global del modelo.

Ventajas:

Capaz de manejar una variedad de problemas, desde la clasificación binaria hasta la clasificación de múltiples clases.

La clasificación supervisada permite obtener modelos con gran precisión si los datos de entrenamiento son representativos.

Se puede aplicar en una amplia gama de industrias, desde el reconocimiento de imágenes hasta la detección de fraudes.

Desventajas:

Requiere de un conjunto de datos etiquetado, lo que puede ser costoso o complicado de obtener.

Los resultados pueden depender en gran medida de la calidad de los datos, como el balance entre clases y la eliminación de ruido.

Algunos algoritmos pueden ser complejos de interpretar, como en el caso de redes neuronales profundas.

La técnica de clasificación es fundamental para resolver problemas donde se necesita asignar etiquetas o categorías a nuevos datos, como la categorización de correos electrónicos, la clasificación de imágenes, o la predicción de comportamientos futuros en función de datos históricos.