Datos Masivos II

Minería de Elementos Frecuentes

Alondra Berzunza

Minería de Elementos Frecuentes

La minería de elementos frecuentes es el proceso de descubrimiento de tendencias o patrones a partir de grandes conjuntos de datos con el objetivo de guiar futuras decisiones.

Surgió en el contexto de los datos de supermercados.

Modelo Mercado-Canasta

Modelo Mercado-Canasta

El “market basket analysis” es una técnica utilizada en el ámbito del análisis de datos y la minería de datos en el campo del comercio minorista y la gestión de ventas.

El objetivo principal de este análisis es descubrir patrones de asociación entre productos que suelen ser comprados juntos por los clientes. Se basa en el concepto de entender qué productos tienden a aparecer en la “cesta de la compra” de un cliente determinado.

Modelo Mercado-Canasta

El MBA es una técnica de minería de datos que permite analizar y establecer combinaciones de productos, identificar los artículos relacionados y qué artículos se compran con frecuencia si es que ya llevamos en el carrito otro artículo.

Propósitos y Aplicaciones

- **Mejorar la disposición de productos en la tienda:** Identificar qué productos suelen comprarse juntos ayuda a los minoristas a organizar y colocar estratégicamente los productos en las tiendas. Pueden ubicar productos complementarios cerca unos de otros para fomentar ventas adicionales.

- **Optimizar estrategias de promoción:** Al entender las asociaciones entre productos, los minoristas pueden diseñar estrategias de promoción más efectivas. Pueden crear ofertas y descuentos que incentiven a los clientes a comprar productos relacionados, aumentando así el valor total de la compra.

- **Personalización y recomendación de productos:** La información obtenida del análisis de la cesta de la compra puede utilizarse para ofrecer recomendaciones personalizadas a los clientes. Las plataformas de comercio electrónico, por ejemplo, pueden sugerir productos relacionados basándose en el historial de compras del cliente.

Propósitos y Aplicaciones

- **Gestión de inventario:** Conocer las asociaciones entre productos ayuda en la gestión de inventario. Los minoristas pueden asegurarse de mantener un inventario adecuado de productos que a menudo se compran juntos, evitando así la falta de existencias y optimizando la cadena de suministro.

- **Segmentación de clientes:** El análisis de la cesta de la compra también puede contribuir a la segmentación de clientes. Identificar patrones de compra específicos puede ayudar a los minoristas a comprender mejor los comportamientos de sus clientes y adaptar estrategias de marketing a grupos específicos.

- **Toma de decisiones estratégicas:** La información derivada del análisis de la cesta de la compra permite a los minoristas tomar decisiones estratégicas informadas sobre la presentación de productos, la estrategia de precios y la planificación de la oferta de productos.

Funcionamiento

1. **Preparación de datos:** Recopilación de transacciones de compra y organización en una matriz.

2. **Cálculo de medidas:** Determinación de la frecuencia de coocurrencia y el soporte para evaluar la importancia de las combinaciones de productos.

3. **Generación de conjuntos frecuentes:** Identificación de conjuntos de productos que superan un umbral de frecuencia usando algoritmos como Apriori o FP-Growth.

Funcionamiento

4. **Generación de reglas de asociación:** Creación de reglas con antecedentes y consecuentes basadas en conjuntos frecuentes, utilizando un umbral de confianza.

5. **Evaluación y selección de reglas:** Evaluación de reglas mediante medidas como la confianza y el lift para seleccionar las más relevantes.

6. **Interpretación de resultados:** Interpretación de las reglas seleccionadas para obtener información valiosa sobre asociaciones y patrones entre los productos comprados por los clientes.

Ejemplo

En los años 90, Wal-Mart descubrió una correlación entre la compra de pañales y cerveza mediante el análisis de datos. Los compradores eran hombres de 25 a 35 años, que adquirían ambos productos los viernes por la tarde. Al colocar cervezas cerca de los pañales, las ventas aumentaron significativamente

Funcionamiento

El algoritmo Apriori es la técnica más común para realizar el análisis de la cesta de mercado.

Se utiliza para La minería de las reglas de asociación , que es un proceso basado en reglas en que se utiliza para identificar correlaciones entre los artículos comprados por los usuarios.

¿Cómo funciona?

Algoritmo A Priori

Algoritmo

El algoritmo Apriori es un algoritmo de machine learning no supervisado que se utiliza para el aprendizaje de reglas de asociación.

El **aprendizaje de reglas de asociación** es una técnica de minería de datos que identifica patrones frecuentes, conexiones y dependencias entre distintos grupos de elementos denominados itemsets en los datos.

Casos de uso más habituales:

- Pronóstico de enfermedades.
- Sistemas de recomendación.
    - Análisis de la cesta de compra.
    - Plataformas de comercio electrónico.

Algoritmo

Introducido en 1994 por Rakesh Agrawal y Ramakrishnan Srikant, el nombre, 'A Priori' reconoce el conocimiento previo de conjuntos de elementos frecuentes. El algoritmo ejecuta iteraciones sobre los datos para identificar k-itemsets, es decir, k elementos que frecuentemente ocurren juntos. Luego utiliza los k-itemsets para identificar los conjuntos de elementos k+1.

Se basa en la idea de que agregar artículos a un grupo comprado con frecuencia solo puede hacerlo menos frecuente, no más; y en la propiedad Apriori que establece que si un conjunto de elementos aparece con frecuencia en un conjunto de datos, todos sus subconjuntos también deben ser frecuentes. Por el contrario, si un conjunto de elementos se identifica como poco frecuente, todos sus superconjuntos se consideran poco frecuentes.

El algoritmo Apriori es aplicable a todo tipo de conjuntos de datos, especialmente los generados por bases de datos transaccionales.

Algoritmo

BENEFICIO PRINCIPAL

Sencillez y adaptabilidad

DESVENTAJA

No son tan eficientes cuando se manejan grandes conjuntos de datos. → alto costo computacional.

Solución: combinar con otras técnicas para mitigar estos problemas.

Funcionamiento

Paso 1: Generación frecuente de conjuntos de elementos.

1.1 Primero se identifican los elementos únicos en el conjunto de datos junto con sus frecuencias.

1.2 Combinar los elementos que aparecen junto con una probabilidad por encima de un umbral especificado en conjuntos de elementos candidatos.

1.3 Filtrar los conjuntos de elementos poco frecuentes para reducir el costo de cómputo en pasos adicionales. → minería de elementos frecuentes.

Funcionamiento

Paso 2: Expandir y luego podar los conjuntos de elementos.

2.1 El algoritmo combina los itemsets frecuentes para formar itemsets más grandes.

2.2 Se podan las combinaciones de itemset más grandes con una probabilidad más baja. Esto reduce aún más el espacio de búsqueda y hace que el cálculo sea más eficiente.

Funcionamiento

Paso 3: Repetir los pasos 1 y 2.

El algoritmo repite los pasos 1 y 2 hasta que todos los conjuntos de elementos frecuentes que cumplen con el umbral de probabilidad definido se generan exhaustivamente. Cada iteración genera asociaciones más complejas y completas en los conjuntos de elementos.

Una vez que A priori creó los conjuntos de elementos, se puede investigar la fuerza de las asociaciones y relaciones generadas.

Medición de conjuntos de elementos - Soporte/Apoyo

El apoyo se define como la relación entre el número de veces que un elemento aparece en las transacciones y el número total de transacciones. Esta métrica define así la probabilidad de que se produzca cada elemento individual en las transacciones. La misma lógica puede extender a los itemsets.

Por ejemplo, en una tienda de venta minorista, 250 de 2000 transacciones realizadas en un día podrían incluir una compra de manzanas.

Con esto, se puede indicar un umbral de soporte mínimo requerido al aplicar el algoritmo Apriori. Esto significa que cualquier elemento o conjunto de elementos con soporte inferior al soporte mínimo especificado se considerará poco frecuente.

Medición de conjuntos de elementos - Confianza

La métrica de confianza identifica la probabilidad de que los elementos o conjuntos de elementos ocurran juntos en los conjuntos de elementos.

Por ejemplo, si hay dos artículos en una transacción, se supone que la existencia de un artículo maneja al otro. El primer elemento o conjunto de elementos es el antecedente y el segundo es el consecuente.

Por lo tanto, la confianza se define como la relación entre el número de transacciones que tienen tanto el antecedente como el consecuente, con el número de transacciones que solo tienen el antecedente.

Si bien la confianza es una buena medida de probabilidad, no es una garantía de una asociación clara entre los elementos. El valor de la confianza puede ser alto por otras razones. Por esta razón, se aplica un umbral de confianza mínimo para filtrar asociaciones débilmente probables durante la minería con reglas de asociación.

Medición de conjuntos de elementos - Lift

El aumento es el factor por el cual la probabilidad de que el elemento A maneje al elemento B es mayor que la probabilidad del elemento A.

Esta métrica cuantifica la fuerza de la asociación entre A y B. Puede ayudar a indicar si existe una relación real entre los elementos en el conjunto de elementos o se agrupan por coincidencia.

El valor de elevación alto indica que la probabilidad de que los elementos se compren juntos es L veces mayor que la de compra de A solo.

El valor de elevación bajo indica que una compra de B que lleva a una compra de A podría ser solo una coincidencia.