JOSÉ AHIRTON BATISTA LOPES FILHO

MACHINE LEARNING ENGINEERING

APRENDIZADO NÃO SUPERVISIONADO

AULA 04

SUMÁRIO

O QUE VEM POR AÍ?

Nesta aula sobre técnicas avançadas de clustering e análise de associação, será abordado o algoritmo Apriori, uma ferramenta fundamental na Mineração de Dados que tem em vista a descoberta de padrões de associação em conjuntos de dados transacionais.

O Apriori é amplamente utilizado de modo a auxiliar na identificação de relações entre itens, principalmente em diferentes transações comerciais, como em carrinhos de compras on-line, registros de vendas em supermercados e históricos de compras de clientes.

Ao analisar esses dados, o algoritmo Apriori é capaz de extrair Regras de Associação, que nada mais são que declarações do tipo "se X, então Y", revelando padrões de compra frequentes e correlações entre diferentes itens.

Essas regras são valiosas para se entender o comportamento do consumidor e otimizar estratégias de marketing e recomendação de produtos, além da melhoria na tomada de decisão empresarial. Durante a aula, serão discutidos exemplos práticos de aplicação do algoritmo Apriori, além dos desafios associados à criação e interpretação das Regras de Associação.

HANDS ON

Neste Hands On, vamos explorar o algoritmo Apriori em uma base de dados de filmes para descobrir associações entre eles, como se estivéssemos analisando um serviço de streaming. Após a preparação dos dados, implementaremos o algoritmo usando a biblioteca Apyori nos utilizando de linguagem Python, definindo parâmetros como suporte e confiança mínimos (mais detalhes nos comentários da codificação).

Na execução do algoritmo, obteremos valores de suporte, confiança e lift para diferentes pares de filmes. Esses resultados nos ajudarão a identificar padrões de comportamento dos usuários, como associações entre filmes frequentemente assistidos juntos, fornecendo insights valiosos que podem ser utilizados para recomendações personalizadas de filmes, além de melhorias na experiência do usuário.

Ao entender quais filmes tendem a ser assistidos em conjunto, podemos aprimorar a oferta de conteúdo, sugerindo filmes relacionados com maior probabilidade de interesse para os usuários, aumentando assim a satisfação e a fidelidade dos clientes.

```
!pip3 install apyori

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from apyori import apriori

!curl -O
https://raw.githubusercontent.com/ahirtonlopes/Unsupervised_L
earning/main/movie_dataset.csv

movie_data = pd.read_csv('movie_dataset.csv', header = None)
num_records = len(movie_data)
print(num_records)

records = []
for i in range(0, num_records):
    records.append([str(movie_data.values[i,j]) for j in
range(0, 20)])

association_rules = apriori(records, min_support=0.004,
min_confidence=0.20, min_lift=3, min_length=2)
association_results = list(association_rules)
```

```
print(len(association_results))

print(association_results[0])

results = []
for item in association_results:

    pair = item[0] #primeiro indice da lista interna
    items = [x for x in pair] #contem item base e adiciona
item

    value0 = str(items[0])
    value1 = str(items[1])

    #indice da lista interna
    value2 = str(item[1])[:7]

    #primeiro indice da lista localizada na posicao 0
    #do tercerio indice da lista interna

    value3 = str(item[2][0][2])[:7]
    value4 = str(item[2][0][3])[:7]

    rows = (value0, value1,value2,value3,value4)
    results.append(rows)

labels = ['Titulo 1','Titulo
2','Support','Confidence','Lift']
movie_suggestion = pd.DataFrame.from_records(results, columns
= labels)

print(movie_suggestion)
```

Código-fonte 1 – Demonstração 6 – Regras de Associação com o Algoritmo Apriori
Fonte: Elaborado pelo autor (2024)

SAIBA MAIS

Na era contemporânea da Ciência de Dados, a necessidade de extração de informações significativas e úteis de conjuntos de dados complexos e volumosos tornou-se imperativa. Duas abordagens fundamentais para lidar com essa necessidade são o Clustering (visto em detalhes anteriormente) e a Análise de Associação. Basicamente, o clustering é uma técnica que visa agrupar dados semelhantes em conjuntos distintos, já a análise de associação procura identificar padrões frequentes ou relações entre diferentes variáveis em um conjunto de dados.

O avanço contínuo da tecnologia e o surgimento de novas fontes de dados geraram conjuntos de dados cada vez maiores e mais complexos. Nesse contexto, as técnicas avançadas de clustering e análise de associação desempenham um papel crucial, permitindo a exploração eficiente e eficaz desses dados massivos. Ao contrário de abordagens mais simples, as técnicas avançadas oferecem maior sofisticação e capacidade de lidar com a complexidade inerente a conjuntos de dados de grande escala.

Nesta jornada de exploração das técnicas avançadas de clustering e análise de associação, nossa intenção é fornecer uma visão aprofundada e cientificamente fundamentada desses métodos. Investigaremos algoritmos sofisticados, teorias subjacentes e aplicações práticas em uma variedade de domínios. Ao fazê-lo, esperamos fornecer insights valiosos e perspectivas inovadoras para enfrentar os desafios cada vez mais complexos e exigentes no campo da ciência e da análise de dados.

Análise de Associação

A análise de associação é uma técnica de mineração de dados que identifica relações significativas entre diferentes variáveis em um conjunto de dados. Ela é frequentemente utilizada em conjuntos de transações de compras, onde o objetivo é descobrir padrões de coocorrência entre itens.

Na análise de associação, as regras de associação são utilizadas para descrever associações frequentes entre itens em um conjunto de transações. Uma regra de associação é geralmente da forma "se X, então Y", indicando que a presença de um conjunto de itens X está associada à presença de outro conjunto de itens Y.

O algoritmo Apriori é um dos algoritmos mais comuns para geração de regras de associação. Ele opera em duas etapas principais: geração de conjuntos de itens frequentes e geração de regras de associação a partir desses conjuntos. Na primeira etapa, o algoritmo identifica todos os conjuntos de itens frequentes com base em um limiar de suporte pré-definido. Na segunda etapa, o algoritmo gera regras de associação a partir dos conjuntos de itens frequentes, calculando métricas como confiança e lift para avaliar a força das regras.

Métricas de Avaliação para Regras de Associação

As métricas mais comuns para avaliar a qualidade das regras de associação incluem suporte, confiança e lift. O suporte mede a frequência absoluta de uma regra em relação ao número total de transações. A confiança mede a proporção de vezes que a regra é verdadeira entre as transações que contêm o antecedente. O lift mede a medida em que o antecedente e o consequente são dependentes entre si, em comparação com a ocorrência esperada deles de forma independente.

Em resumo, a análise de associação é uma técnica poderosa para descobrir padrões interessantes em grandes conjuntos de dados transacionais, como registros de compras. O algoritmo Apriori é uma abordagem comum para a geração de regras de associação e a escolha das métricas de avaliação adequadas é fundamental para identificar as associações mais relevantes e significativas.

Aplicações e Casos de Uso Mais Relevantes

Nesta seção, exploraremos diversos estudos de caso que demonstram a aplicação prática de técnicas avançadas de clustering e análise de associação em cenários do mundo real.

Segmentação de Clientes em um Banco usando Clustering Hierárquico

Imagine que um banco deseja segmentar sua base de clientes para oferecer serviços personalizados. Por exemplo, se utilizando de clustering hierárquico, é possível agrupar os clientes com base em características semelhantes, como idade, renda, histórico de transações e preferências bancárias. Isso permite que o banco identifique grupos de clientes com necessidades específicas e adapte suas estratégias de marketing e atendimento para atender melhor a cada segmento.

Detecção de Fraudes em Transações Financeiras usando DBSCAN

Uma instituição financeira busca identificar atividades fraudulentas em sua rede de transações. O algoritmo DBSCAN pode ser aplicado de modo a auxiliar na detecção de agrupamentos de transações incomuns que se desviam do comportamento típico dos clientes. Ao identificar esses agrupamentos como potenciais casos de fraude, a instituição pode tomar medidas preventivas para mitigar o impacto financeiro e proteger seus clientes.

Análise de Cestas de Compras usando o Algoritmo Apriori

Um supermercado deseja entender os padrões de compra dos clientes para otimizar o layout da loja e as estratégias de marketing. Por meio da análise de cestas de compras com o algoritmo Apriori, o supermercado pode identificar associações frequentes entre produtos comprados juntos. Isso possibilita a criação de recomendações de produtos personalizadas, promoções cruzadas e ajustes na disposição dos itens nas prateleiras para aumentar as vendas e melhorar a experiência do cliente.

Esses estudos de caso ilustram como as técnicas avançadas de clustering e análise de associação podem ser aplicadas em diversos setores para resolver problemas específicos e impulsionar a tomada de decisões estratégicas. Ao compreender os padrões subjacentes nos dados, as organizações podem ganhar insights valiosos que as ajudam a se destacar em um mercado cada vez mais competitivo.

Ao aplicar técnicas avançadas de clustering e análise de associação em projetos do mundo real, é importante estar ciente dos desafios comuns e considerações práticas que podem surgir ao longo do processo. Além do nosso Hands On, teremos uma discussão a respeito desses desafios e recomendação de diretrizes para a implementação bem-sucedida dessas técnicas.

Livros

"Pattern Recognition and Machine Learning" de Christopher M. Bishop: embora este livro seja mais conhecido por abordar tópicos de reconhecimento de padrões e aprendizado de máquina supervisionado, ele também inclui uma discussão sobre técnicas de aprendizado não supervisionado, como clustering e redução de dimensionalidade, que são relevantes para a análise de associação.

"Data Mining: Concepts and Techniques" de Jiawei Han e Micheline Kamber: este livro é uma referência amplamente utilizada na área de mineração de dados e cobre diversos aspectos da análise de associação, incluindo algoritmos como Apriori e FP-Growth.

"Association Rules Mining: A Recent Overview" de Thanuja K.M. e Nirmala S.: este artigo oferece uma visão geral atualizada sobre a análise de associação, abordando diferentes algoritmos, métricas de avaliação e aplicações.

Artigos Científicos

“Fast algorithms for mining association rules” de Rakesh Agrawal e Srikant Ramakrishnan: este é um dos papers pioneiros sobre o algoritmo Apriori, amplamente utilizado para mineração de regras de associação em grandes conjuntos de dados.

“Mining the most interesting rules” de Roberto J. Bayardo Jr e Rakesh Agrawal: este paper propõe uma abordagem para identificar as regras de associação mais interessantes e relevantes, levando em consideração critérios como suporte e confiança.

Essas obras demonstram algumas das técnicas seminais e desafios na área de construção de regras de associação. Desde o pioneiro trabalho de Agrawal e Srikant em 1994, que introduziu o algoritmo Apriori, até os mais recentes desenvolvimentos em mineração de regras de associação, esses estudos têm sido fundamentais para o avanço do campo. A mineração de regras de associação desempenha um papel crucial no campo do Aprendizado de Máquina, fornecendo insights valiosos sobre padrões e relações nos dados.

Ao identificar associações entre diferentes itens ou variáveis, essas técnicas permitem que os sistemas de aprendizado façam previsões mais precisas e tomem decisões mais informadas. Além disso, a mineração de regras de associação é amplamente utilizada em áreas como análise de mercado, recomendação de produtos e detecção de fraudes, demonstrando sua importância em uma variedade de aplicações do mundo real. Portanto, o estudo e a compreensão dessas técnicas são essenciais para qualquer praticante ou pesquisador(a) interessado(a) em explorar todo o potencial do Aprendizado de Máquina.

Algoritmo Apriori: Uma Abordagem Técnica

O algoritmo Apriori é uma técnica bastante popular na área de Mineração de Regras de Associação, frequentemente utilizada para a descoberta de padrões de compra em conjuntos de dados transacionais. Nesta seção, vamos explorar detalhadamente o funcionamento do algoritmo, incluindo suas etapas e a lógica por trás de sua implementação.

O algoritmo Apriori baseia-se no princípio de Apriori, que sugere que se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser. Isso significa que se um conjunto de itens não é frequente, então seus superconjuntos também não serão. Esse princípio é crucial para reduzir o espaço de busca durante a geração de regras de associação.

Etapas do Algoritmo

O algoritmo Apriori consiste em algumas etapas distintas.

- **Geração de Conjuntos de Itens Frequentes**

Inicialmente, o algoritmo identifica todos os itens únicos no conjunto de dados e calcula sua frequência de ocorrência;

Em seguida, ele gera todos os conjuntos de itens de tamanho 1 (itemsets de 1 item) que atendem ao critério de suporte mínimo.

- **Geração de Candidatos**

Com base nos itemsets frequentes de tamanho k-1, o algoritmo gera candidatos para itemsets de tamanho k;

Esses candidatos são formados combinando os itemsets frequentes de tamanho k-1 de maneiras específicas.

- **Poda (Pruning)**

Durante a geração de candidatos, o algoritmo utiliza o princípio de Apriori para podar os conjuntos de itens que não podem ser frequentes com base nos conjuntos de tamanho k-1.

- **Verificação de Suporte**

Após a geração de candidatos, o algoritmo realiza uma passagem adicional pelo conjunto de dados para verificar o suporte de cada candidato. Apenas os candidatos que atendem ao critério de suporte mínimo são considerados frequentes.

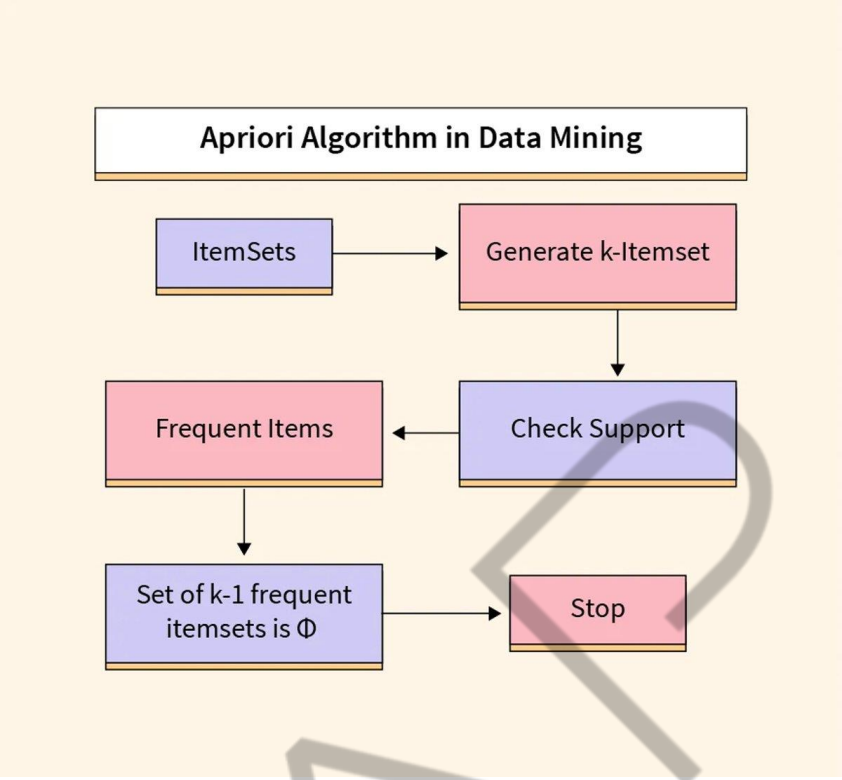

Figura 1 – Visão do processo de construção de Regras de Associação via Algoritmo Apriori
Fonte: Utkarsh (2024)

Desafios Comuns na Implementação do Apriori

Seleção de Parâmetros: a escolha adequada de parâmetros, como o número de clusters ou o limiar de suporte na análise de associação, pode influenciar significativamente os resultados. A determinação desses parâmetros pode ser uma tarefa desafiadora e requer experimentação cuidadosa.

Interpretação dos Resultados: interpretar e validar os resultados do clustering e da análise de associação pode ser complexo, especialmente em conjuntos de dados grandes e de alta dimensionalidade. Identificar padrões relevantes e distinguir entre agrupamentos significativos e ruído pode exigir expertise especializada.

Escalabilidade: lidar com conjuntos de dados grandes e complexos pode representar um desafio em termos de escalabilidade computacional. Algoritmos de clustering e análise de associação eficientes são necessários para lidar com grandes volumes de dados de maneira eficaz.

Considerações Práticas e Futuro em Regras de Associação

Preparação de Dados: uma etapa crucial é a preparação adequada dos dados, incluindo limpeza, normalização e seleção de recursos relevantes. Dados de alta

qualidade e bem estruturados são essenciais para obter resultados precisos e significativos.

Avaliação de Resultados: é importante realizar uma avaliação rigorosa dos resultados obtidos, utilizando métricas apropriadas para medir a qualidade dos agrupamentos ou a relevância das regras de associação geradas. A validação cruzada e outras técnicas de validação podem ser úteis para garantir a robustez dos resultados.

Iteração e Ajuste: o processo de clustering e análise de associação muitas vezes envolve iteração e ajuste dos algoritmos e parâmetros. É importante estar preparado(a) para experimentar diferentes abordagens e refinamentos para alcançar os melhores resultados possíveis.

Ao enfrentar esses desafios e considerações práticas, as organizações podem maximizar o potencial das técnicas avançadas de clustering e análise de associação para extrair insights valiosos dos dados e impulsionar a tomada de decisões informadas. Com uma abordagem cuidadosa e metodológica, é possível superar os obstáculos e obter benefícios significativos para o negócio.

Nesta aula, exploramos uma variedade de técnicas avançadas de clustering e análise de associação, desde algoritmos específicos até estudos de caso de aplicação em cenários do mundo real. Recapitulamos os principais pontos discutidos e delineamos algumas perspectivas interessantes para futuras pesquisas e desenvolvimentos nessas áreas. À medida que avançamos para o futuro, há várias áreas promissoras para pesquisas e desenvolvimentos adicionais em clustering e análise de associação. Alguns dos tópicos incluem:

- **Desenvolvimento de Algoritmos Mais Eficientes**: pesquisas contínuas são necessárias para o desenvolvimento de algoritmos de clustering e análise de associação mais eficientes e escaláveis, capazes de lidar com conjuntos de dados cada vez maiores e mais complexos.
- **Integração com Aprendizado Profundo**: a integração de técnicas de clustering e análise de associação com técnicas de Aprendizado Profundo oferece oportunidades emocionantes para análise de dados em uma variedade de domínios, incluindo Visão Computacional, Processamento de Linguagem Natural e Biologia Computacional.

- **Aplicações em Tempo Real**: os últimos desenvolvimentos e pesquisas têm se concentrado na evolução de técnicas de clustering e análise de associação que possam ser aplicadas em tempo real, permitindo uma análise contínua e em tempo real de grandes fluxos de dados.
- **Interpretabilidade e Explicabilidade**: melhorar a interpretabilidade e explicabilidade dos resultados do clustering e análise de associação é um desafio importante. Pesquisas adicionais podem se concentrar no desenvolvimento de métodos e técnicas para tornar os resultados mais compreensíveis e utilizáveis para os usuários finais. À medida que avançamos, é essencial continuar explorando e inovando nessas áreas, aproveitando todo o potencial das técnicas avançadas de clustering e análise de associação para extrair insights valiosos e impulsionar o progresso em uma ampla gama de aplicações práticas.

MERCADO, CASES E TENDÊNCIAS

Artigo EN-US – Mineração de Regras de Associação no Auxílio a Saúde Pública

O artigo "Mineração de regras de associação para malária - Uma análise da malária na Amazônia Legal brasileira usando regras de associação divergentes" aborda uma aplicação crucial de técnicas de mineração de dados na área da saúde pública. A malária é uma doença grave que representa um desafio significativo em regiões como a Amazônia, em que fatores ambientais e socioeconômicos podem influenciar sua propagação.

Neste estudo, pesquisadores e pesquisadoras utilizaram métodos avançados de mineração de regras de associação para analisar dados epidemiológicos da malária na Amazônia Legal brasileira. Em vez de focar apenas em associações positivas entre variáveis, como é comum em análises de regras de associação, os(as) autores(as) exploraram associações divergentes, que destacam padrões negativos ou inesperados nos dados. Leia mais aqui.

O QUE VOCÊ VIU NESTA AULA?

Nesta aula exploramos os fundamentos da Mineração de Regras de Associação e o Algoritmo Apriori. A Mineração de Regras de Associação é uma técnica essencial em Aprendizado de Máquina, a qual nos permite descobrir padrões interessantes e relações ocultas entre itens em grandes conjuntos de dados. O algoritmo Apriori, desenvolvido por Agrawal e Srikant em 1994, é um dos métodos mais populares para realizar essa tarefa.

Durante as seções do presente conteúdo, aprendemos como preparar os dados e aplicar o algoritmo Apriori para identificar regras de associação significativas. Exploramos conceitos-chave, como suporte, confiança e lift, que nos ajudam a avaliar a importância e a relevância das regras descobertas a partir de um dataset de filmes.

Além disso, discutimos alguma das estratégias usadas de modo a ajustar os parâmetros do algoritmo, como o suporte mínimo e a confiança mínima, para obter resultados mais úteis e precisos.

Ao aplicar o Apriori em diferentes conjuntos de dados, ganhamos insights valiosos sobre padrões de comportamento e relações entre itens. Esses insights têm aplicações em uma variedade de domínios, desde recomendação de produtos em comércio eletrônico até análise de mercado de ações e detecção de fraudes.

Compreender os conceitos e técnicas da Mineração de Regras de Associação e do algoritmo Apriori nos permite extrair conhecimento significativo de dados complexos e tomar decisões mais informadas em uma ampla gama de contextos.

REFERÊNCIAS

AGRAWAL, R.; RAMAKRISHNAN, S. **Fast algorithms for mining association rules**. In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB Endowment, 1994. Disponível em: <https://www.vldb.org/conf/1994/P487.PDF>. Acesso em: 12 jul. 2024.

BAYARDO JR, R. J.; AGRAWAL, R. **Mining the most interesting rules**. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999. Disponível em: <https://dl.acm.org/doi/10.1145/312129.312219>. Acesso em: 12 jul. 2024.

BISHOP, C. M. **Pattern Recognition and Machine Learning.** New York: Springer, 2006. Disponível em: <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>. Acesso em: 12 jul. 2024.

HAN, J.; KAMBER, M. **Data Mining**: Concepts and Techniques. San Francisco: Morgan Kaufmann, 2006. Disponível em: <https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>. Acesso em: 12 jul. 2024.

KOTSIANTIS, S.; KANELLOPOULOS, D. **Association Rules Mining**: A Recent Overview. 2006. Disponível em: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=73a19026fb8a6ef5bf238ff472f31100c33753d0>. Acesso em: 12 jul. 2024.

UTKARSH. **Frequent Patterns in Data Mining**. 2024. Disponível em: <https://www.scaler.com/topics/data-mining-tutorial/frequent-pattern-mining/>. Acesso em: 12 jul. 2024.

PALAVRAS-CHAVE

Inteligência Artificial. Aprendizado de Máquina. Aprendizado Não Supervisionado.