Projeto 1
Teoria e Aplicação de Grafos

João Vítor Maia - 190110007
Mateus Lucas Oliveira Filho - 221000080

[1]Dep. Ciência da Computação – Universidade de Brasília (UnB)
CIC0135 - INTRODUCAO A INTELIGENCIA ARTIFICIAL
27 de maio de 2025

1. Introdução

O crescimento do comércio eletrônico e a disponibilidade cada vez maior de dados de transações de clientes tornaram os sistemas de recomendação uma ferramenta essencial para personalizar a experiência de compra e aumentar a satisfação do usuário. Em particular, no contexto do agronegócio, produtores rurais e distribuidores dispõem de grande variedade de produtos agrícolas, mas muitas vezes têm dificuldades em identificar quais itens são mais relevantes para diferentes perfis de clientes. Este trabalho apresenta a implementação de um sistema de recomendação baseado no algoritmo K-Nearest Neighbors (KNN), visando sugerir produtos agrícolas personalizados a partir do histórico de compras de clientes.

Para isso, foram utilizados dois conjuntos de dados principais: (i) as tabelas de produção agrícola por Região Administrativa (RA) do Distrito Federal, convertidas para o formato CSV por meio dos scripts `extract.py` e `format.py`; e (ii) um conjunto sintético de 100 vendas geradas para 50 clientes, criado pelo script `fake_customers_generation.py`, que incorpora comportamentos probabilísticos de reaquisição (70% de chance para produtos já comprados, 25% para produtos produzidos na mesma RA do cliente e 5% para demais produtos). Esses dados permitem avaliar a capacidade do modelo de capturar padrões de preferência e de recomendar itens de forma coerente.

O restante deste documento está organizado da seguinte forma: na Seção "Coleta de Dados" detalham-se os procedimentos de extração e formatação dos dados; em "Sistema de Recomendação Baseado em KNN" descrevem-se o pré-processamento, o modelo, o processo de recomendação e a avaliação; por fim, a seção de Conclusão apresenta um resumo dos resultados, as principais limitações identificadas e sugestões de melhorias para trabalhos futuros.

Coleta de Dados

Para coleta de dados utilizamos as tabelas de produções agrícolas por RA fornecidas via moodle, utilizamos os arquivos extract.py e format.py para a conversão dessas tabelas em csv, possibilitando assim melhor manejamento dos dados para o treinamento da IA. Para dados de venda utilizamos o script fake_customers_generation.py para gerar 100 vendas de 50 clientes, adicionamos alguns comportamentos na geração dos dados, para que verifiquemos se os mesmos se repetem no resultado do treinamento, sendo eles: produtos já comprados anteriormente por um cliente têm 70% mais chance de serem comprados novamente, produtos produzidos pela mesma RA do cliente, possuem 25% mais chance

de serem comprados novamente, o restante dos produtos possuem 5% de chance de serem comprados.

2. Sistema de Recomendação Baseado em KNN

O sistema de recomendação implementado utiliza o algoritmo K-Nearest Neighbors (KNN) para identificar clientes com padrões de compra similares e, a partir deles, gerar recomendações personalizadas de produtos agrícolas. O processo de treinamento e recomendação é realizado em várias etapas, descritas a seguir.

2.1. Pré-processamento dos Dados

Inicialmente, os dados de vendas são carregados de um arquivo CSV contendo as transações de compra. Cada transação inclui:

- Cliente que realizou a compra
- Produto adquirido
- Quantidade comprada
- Localidade da compra

Estes dados são transformados em uma matriz cliente-produto (matriz de utilidade) usando uma tabela pivot, onde:

- Cada linha representa um cliente
- Cada coluna representa um produto
- Cada célula contém a quantidade total do produto comprada pelo cliente

Para otimizar o uso de memória e processamento, a matriz é convertida para o formato esparso usando `csr_matrix` da biblioteca SciPy, já que muitos clientes não compram todos os produtos disponíveis.

2.2. Modelo KNN

O modelo KNN é implementado utilizando a classe `NearestNeighbors` do scikit-learn com os seguintes parâmetros:

- `n_neighbors = 5`: Número de vizinhos mais próximos a considerar
- `metric = 'cosine'`: Similaridade do cosseno como métrica de distância
- `algorithm = 'brute'`: Busca por força bruta, adequada para datasets pequenos/médios

A similaridade do cosseno foi escolhida por ser particularmente eficaz para dados esparsos, pois:

- Considera apenas produtos comprados em comum
- É invariante à escala das quantidades compradas
- Captura bem o padrão de preferências dos clientes

2.3. Processo de Recomendação

Para gerar recomendações para um cliente específico, o sistema:

1. Localiza o cliente na matriz de utilidade
2. Encontra os K clientes mais similares usando KNN
3. Remove o próprio cliente da lista de vizinhos
4. Soma as quantidades compradas pelos vizinhos para cada produto
5. Seleciona os N produtos mais frequentes como recomendações

2.4. Avaliação do Modelo

O sistema é avaliado usando validação hold-out, onde:

- 10% das compras de cada cliente são separadas para teste
- O modelo é treinado nos 90% restantes
- São calculadas métricas de precisão e recall para as recomendações

As métricas são definidas como:

$$\text{Precisão@K} = \frac{\text{número de recomendações relevantes}}{\text{número total de recomendações}} \tag{1}$$

$$\text{Recall@K} = \frac{\text{número de recomendações relevantes}}{\text{número de itens relevantes}} \tag{2}$$

Os resultados mostram que o sistema consegue um bom equilíbrio entre precisão e recall, com valores típicos de:

- Precisão@10: aproximadamente 0.22
- Recall@10: aproximadamente 0.38

2.5. Características do Sistema

O sistema implementado possui algumas características importantes:

- **Escalabilidade**: O uso de matrizes esparsas permite lidar com grandes volumes de dados
- **Personalização**: As recomendações são baseadas no histórico individual de cada cliente
- **Interpretabilidade**: O processo de recomendação é transparente e fácil de entender
- **Flexibilidade**: Os parâmetros K (vizinhos) e N (recomendações) podem ser ajustados

2.6. Limitações e Possíveis Melhorias

O sistema atual apresenta algumas limitações que poderiam ser endereçadas em trabalhos futuros:

- Não considera a temporalidade das compras
- Não leva em conta a sazonalidade dos produtos agrícolas
- Não incorpora informações sobre a localidade do cliente
- Não considera o feedback explícito dos clientes

Possíveis melhorias incluem:

- Incorporar pesos temporais para dar mais importância a compras recentes
- Adicionar informações de sazonalidade dos produtos
- Implementar um sistema híbrido com filtragem baseada em conteúdo
- Incluir feedback explícito dos clientes quando disponível

Conclusão

Neste projeto, implementamos um sistema de recomendação colaborativa baseado em KNN que demonstrou equilíbrio entre precisão e recall, com valores típicos de Precisão@10 em aproximadamente 0,22 e Recall@10 em torno de 0,38. A utilização de matrizes esparsas garantiu escalabilidade para cenários com grande número de clientes e produtos, enquanto a métrica de similaridade de cosseno mostrou-se eficaz para capturar padrões de compra mesmo em ambientes altamente esparsos.

Entretanto, identificamos algumas limitações que podem ser abordadas em trabalhos futuros: (i) a ausência de tratamento da temporalidade das transações, que poderia priorizar recomendações mais recentes; (ii) a falta de incorporação de sazonalidade, importante para produtos agrícolas; (iii) a não utilização de informações explícitas de feedback dos clientes; e (iv) a ausência de características de localização geográfica no modelo. Sugerimos, portanto, explorar algoritmos híbridos que combinem filtragem colaborativa e baseada em conteúdo, bem como incorporar pesos temporais e dados de sazonalidade para aprimorar a qualidade das recomendações.

Bibliografia

O trabalho pode ser encontrado no notebook.

Referências

[1] Hagberg, A., Schult, D., Swart, P. *Exploring network structure, dynamics, and function using NetworkX*. 2008.

[2] Leskovec, J., Krevl, A. *SNAP Datasets: Stanford Large Network Dataset Collection*. 2014.