Avaliação de Algoritmos de Clusterização para Agrupamento de Descrições de Produtos em Notas Fiscais Eletrônicas

Jonas Gabriel L. de Araújo[1], **Thaís G. do Rêgo**[1], **Yuri de A. M. Barbosa**[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa - PB - Brasil

jonas.araujo@academico.ufpb.br, gaudenciothais@gmail.com, yuri@ci.ufpb.br

Abstract. *The electronic invoice is essential for the tax audit process. This paper evaluates the effectiveness of clustering algorithms in grouping product descriptions from electronic invoices, a key document in tax audits. Due to the lack of standardization in these descriptions, clustering becomes a challenge. Using string similarity and adjustments for different units of measurement, DBSCAN, HDBSCAN, OPTICS, and Agglomerative Clustering were tested. Evaluation metrics included the Silhouette Coefficient, Calinski-Harabasz Index, and the percentage of grouped products. HDBSCAN showed the best initial performance, and the subclustering stage, while improving metrics, introduced inconsistencies in the groups.*

Resumo. *A nota fiscal eletrônica é essencial para o processo de auditoria fiscal. Este artigo avalia a eficácia de algoritmos de clusterização para agrupar descrições de produtos em notas fiscais eletrônicas, um desafio devido à falta de padronização nos registros. Usando similaridade de strings e ajustes para unidades de medida, foram testados DBSCAN, HDBSCAN, OPTICS e Agglomerative Clustering. As métricas de avaliação incluíram o Coeficiente de Silhueta, Índice de Calinski-Harabasz e a porcentagem de produtos agrupados. O HDBSCAN apresentou o melhor desempenho inicial, e a subclusterização, apesar de melhorar as métricas, introduziu inconsistências nos agrupamentos.*

1. Introdução

As Notas Fiscais Eletrônicas (NF-e) são um marco na modernização dos processos fiscais no Brasil, ao melhorar o controle e a fiscalização tributária, o que já resultou em avanços na arrecadação de impostos e no processo de auditoria [Vieira et al. 2019, Neto and Lopo Martinez 2016]. No entanto, a análise dessas notas enfrenta desafios devido à falta de padronização nas descrições de produtos, com erros ortográficos, abreviações e variações nas unidades de medida [Mazzarolo et al. 2022]. Essa inconsistência dificulta a organização e comparação de dados, exigindo técnicas computacionais para agrupar descrições similares e auxiliar na auditoria fiscal, que requer a correspondência entre o inventário das empresas e as notas emitidas por elas. Dessa forma, o uso de algoritmos de agrupamento facilita a fiscalização e melhora a eficiência do processo [Ribeiro et al. 2018].

Neste contexto, este estudo busca avaliar algoritmos de agrupamento, como DBSCAN [Ester et al. 1996], HDBSCAN [Campello et al. 2013], OPTICS [Ankerst et al. 1999] e *Agglomerative Clustering* (AGG) [Steinbach et al. 2000], para

agrupar descrições de produtos e identificar quais algoritmos oferecem o melhor desempenho na organização e interpretação dos dados. Para isso, foi empregada uma métrica personalizada no cálculo da matriz de distâncias, baseada em similaridade entre *strings* e a análise de uma segunda etapa de agrupamento dos dados.

2. Trabalhos relacionados

Nesta seção, serão abordados alguns trabalhos que contribuíram para tentar resolver as diferenças na padronização nas descrições dos produtos, a fim de melhorar o processo de fiscalização tributária no Brasil [Mazzarolo et al. 2022].

O trabalho de [Schulte et al. 2022] apresentou o ELINAC, um modelo que combina *autoencoder* e busca binária para agrupar descrições de produtos em notas fiscais. O método filtra as descrições, considerando apenas o nome e informações numéricas, como quantidade e dosagem. Embora eficiente, ele tem limitações ao distinguir produtos com variações sutis, como sabor.

A revisão de [Ahmed et al. 2022] aponta que a representação vetorial de textos curtos é desafiadora devido à alta dimensionalidade e ao ruído. O estudo de [Marinho et al. 2024] comparou representações textuais para classificar inconsistências em notas fiscais, calculando a similaridade entre a descrição do produto e a oficial da Nomenclatura Comum do Mercosul (NCM). Concluiu-se que a distância de edição de *strings* teve melhor desempenho preditivo do que *embeddings*, apesar de não considerar a similaridade entre produtos.

Este estudo se diferencia ao focar na avaliação de algoritmos de agrupamento e na representação de descrições de NF-es utilizando similaridade de *strings*. Enquanto outros trabalhos abordam redes neurais, detecção de fraudes e visualização de dados, este estudo explora a eficácia dos algoritmos de clusterização para organizar e interpretar as descrições de produtos em notas fiscais.

3. Metodologia

Esta seção descreve a base de dados, o cálculo da matriz de distâncias e os algoritmos de clusterização utilizados.

3.1. Base de dados

Foram usadas duas bases: uma base sintética com 22 descrições, contendo ruídos típicos [Mazzarolo et al. 2022], e uma base real cedida pela Secretaria da Fazenda da Paraíba (SEFAZ-PB) com 507 descrições. As descrições foram normalizadas, removendo caracteres especiais e convertendo tudo para caracteres maiúsculos.

3.2. Matriz de distâncias

Uma matriz de distâncias é uma matriz quadrada que contém as distâncias entre todos os pares de elementos do banco de dados. Neste trabalho, a matriz foi feita a partir de uma métrica personalizada, baseada na similaridade de Jaro [Jaro 1989]. O valor da similaridade varia entre 0 e 1, onde 0 indica que as *strings* não têm correspondências e 1 indica que as *strings* são idênticas. Entretanto, para o conceito de distância, quanto mais próximo de 0, mais próximos são dois pontos. Dessa forma, para computar a matriz de distância, foi calculado o complemento da similaridade de Jaro, ou seja, $1-\text{JaroSimilarity}$.

Além da similaridade textual, foi introduzido um cálculo adicional para diferenciar produtos com o mesmo nome, mas com medidas distintas, como “200 ML” e “10 KG”. Isso evita que produtos com variação apenas na quantidade sejam considerados iguais. Para implementar esse ajuste, as medidas foram extraídas por meio da expressão regular 1 [Lucena et al. 2022], e convertidas em mililitros, gramas ou metros. Quando as medidas diferem, adiciona-se uma penalidade de 0,3 ao complemento da similaridade de Jaro, valor que foi escolhido após testes com variações entre 0,1 e 0,5.

$$(?:\backslash d*[,]?\backslash d+?\backslash s?(?:kg|ml|mm|l|lt|gr|grs|g|metros|m|gb|k|cm|mg)\backslash b) \quad (1)$$

3.3. Algoritmos de *Clusterização*

Para o agrupamento, este estudo avaliou 4 algoritmos diferentes: DBSCAN [Ester et al. 1996], HDBSCAN [Campello et al. 2013], OPTICS [Ankerst et al. 1999] e AGG [Steinbach et al. 2000]. Todos os algoritmos usados foram implementados pela biblioteca *scikit-learn*, versão 1.5.1, e nenhuma métrica de distância foi passada para os algoritmos, uma vez que a matriz já está pré-computada.

Os algoritmos foram usados em duas etapas: o agrupamento inicial e a *subclusterização* dos grupos de *outliers*, aplicada apenas na base real. Para o agrupamento inicial, foi definida uma distância máxima de agrupamento de $0{,}1$ e um tamanho mínimo de *cluster* sendo igual a 2. Para a segunda etapa, a distância foi igual a $0{,}2$. Os parâmetros de distância foram escolhidos após avaliação do agrupamento com variações entre $0{,}05$ e $0{,}2$ e os demais hiperparâmetros possuem os valores padrões da biblioteca.

Para avaliar o resultado dos agrupamentos, foram utilizadas duas métricas principais: o Coeficiente de Silhueta [Rousseeuw 1987], que avalia a coesão dos *clusters*, e o Índice de Calinski-Harabasz (CH) [Caliński and JA 1974], que mede a separação entre os grupos. O cálculo dessas métricas foi feito utilizando as distâncias entre pontos pré-computadas. Além disso, foi considerada a porcentagem de produtos agrupados para avaliar a cobertura dos dados pelos algoritmos de agrupamento.

4. Resultados e discussões

A Tabela 1 apresenta os resultados da primeira etapa dos experimentos. É importante ressaltar que o algoritmo AGG não gera um grupo de *outliers* identificado como -1, o que exigiu um ajuste no cálculo das métricas para esse caso. Especificamente, todos os grupos individuais, que contêm apenas um produto, foram considerados como pertencentes ao grupo -1, permitindo que as métricas fossem calculadas de forma consistente.

Na base sintética, DBSCAN, OPTICS e AGG produziram *clusters* idênticos, enquanto o HDBSCAN teve desempenho superior, distinguindo produtos com variações de sabor, mas não separando bem produtos de medidas diferentes. Nos dados reais, o HDBSCAN obteve as melhores métricas gerais, enquanto o OPTICS teve o maior coeficiente de Silhueta, mas o menor índice de CH, sugerindo que seus *clusters* não estavam bem separados.

A Tabela 2 apresenta os resultados da *subclusterização* dos grupos de produtos considerados *outliers* na base de dados real. Todas as métricas possuíram aumentos nos valores, quando comparados à primeira clusterização, especialmente na utilização do HDBSCAN, tanto na primeira, quanto na segunda etapa.

Tabela 1. Avaliação dos algoritmos de clusterização no agrupamento inicial

Base de Dados	Algoritmo	Silhueta	CH	Produtos agrupados (%)
Base Controlada	DBSCAN	0,490	7,97	98,16
	HDBSCAN	0,563	15,58	99,80
	OPTICS	0,490	7,97	98,16
	AGG	0,490	7,97	98,16
Base SEFAZ-PB	DBSCAN	0,686	21,71	86,19
	HDBSCAN	0,726	43,40	94,08
	OPTICS	0,730	17,98	85,99
	AGG	0,696	20,18	75,79

Embora as métricas tenham melhorado com a segunda etapa de clusterização usando o HDBSCAN, surgiram inconsistências nos agrupamentos. Por exemplo, produtos como "BOM TRIGO PREP. EMULSIF." e "MARG. MEDALHA DE OURO" foram agrupados erroneamente no mesmo *cluster*. Isso indica que a fase adicional pode priorizar a melhoria das métricas, mas comprometer a consistência semântica, tornando os *clusters* menos úteis ou interpretáveis na prática.

Tabela 2. Avaliação dos algoritmos de clusterização no segundo agrupamento

Primeira Etapa	Segunda Etapa	Silhueta	CH	Produtos agrupados (%)
DBSCAN	DBSCAN	0,718	29,02	90,13
	HDBSCAN	0,737	79,30	97,63
	OPTICS	0,717	27,50	89,74
	AGG	0,719	28,06	89,94
HDBSCAN	DBSCAN	0,729	46,28	94,47
	HDBSCAN	0,740	125,55	99,21
	OPTICS	0.729	46,28	94,47
	AGG	0,729	46,28	94,47
OPTICS	DBSCAN	0,761	25,25	89,94
	HDBSCAN	0,779	73,11	97,63
	OPTICS	0,760	22,96	89,54
	AGG	0,763	23,46	89,74
AGG	DBSCAN	0,728	27,42	89,94
	HDBSCAN	0,746	75.52	97,43
	OPTICS	0,727	26.00	89,54
	AGG	0,729	26,53	89,74

5. Considerações finais

Este estudo avaliou os algoritmos de clusterização DBSCAN, HDBSCAN, OPTICS e AGG para agrupar descrições de produtos em NF-e, utilizando similaridade de *strings* como representação de dados. O HDBSCAN apresentou o melhor desempenho inicial, mas a segunda etapa de agrupamento gerou inconsistências. DBSCAN e OPTICS tiveram métricas um pouco inferiores, porém com menos irregularidades. Sugere-se, como trabalhos futuros, testar o método em bases maiores e explorar representações como *embeddings* e redes neurais para padronização.

Referências

Ahmed, M., Tiun, S., Omar, N., and Sani, N. S. (2022). Short text clustering algorithms, application and challenges: A survey. *Applied Sciences*.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60.

Caliński, T. and JA, H. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27.

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pei, J., Tseng, V. S., Cao, L., Motoda, H., and Xu, G., editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Lucena, L. F., de Menezes e Silva Filho, T., do Rêgo, T. G., and Malheiros, Y. (2022). Automatic recognition of units of measurement in product descriptions from tax invoices using neural networks. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 156–165, Cham. Springer International Publishing.

Marinho, M., Weigang, L., Oliveira, V., and Borges, V. (2024). Estratégias computacionais baseadas em similaridade de textos e visualização exploratória para a identificação de inconsistências em notas fiscais eletrônicas.

Mazzarolo, J., Steinmetz, R., and Mergen, S. (2022). Um estudo sobre a falta de padronização na descrição de produtos em notas fiscais eletrônicas. In *Anais da XVII Escola Regional de Banco de Dados*, pages 31–40, Porto Alegre, RS, Brasil. SBC.

Neto, H. and Lopo Martinez, A. (2016). Nota fiscal de serviÇos eletrÔnica: Uma anÁlise dos impactos na arrecadaÇÃo em municÍpios brasileiros. *Revista de Contabilidade e Organizações*, 10:49.

Ribeiro, L., Brandão, W., Marques, I., Andrade, P., Júnior, R., Oliveira, F., and Kelles, R. (2018). Reconhecimento de entidades nomeadas em itens de produto da nota fiscal eletrônica. 36:116–126.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Schulte, J. P., Giuntini, F. T., Nobre, R. A., Nascimento, K. C. d., Meneguette, R. I., Li, W., Gonçalves, V. P., and Rocha Filho, G. P. (2022). Elinac: Autoencoder approach for electronic invoices data clustering. *Applied Sciences*, 12(6).

Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques.

Vieira, P. A., Pimenta, D. P., Cruz, A. F. d., and Souza, E. M. S. d. (2019). Efeitos do programa de nota fiscal eletrônica sobre o aumento da arrecadação do estado. *Revista de Administração Pública*, 53(2):481–491.