Algoritmos de agrupamento aplicados à detecção de fraudes

Gabriel Covello Furlanetto[1], Veronica Oliveira de Carvalho[1], Alexandro Baldassin[1], Aleardo Manacero[2]

[1]Departamento de Estatística, Matemática Aplicada e Ciência da Computação – Universidade Estadual Paulista (UNESP) – Rio Claro – SP – Brasil

[2]Departamento de Ciência da Computação e Estatística – Universidade Estadual Paulista (UNESP) – São José do Rio Preto – SP – Brasil

{gabriel.furlanetto,veronica.carvalho}@unesp.br

{aleardo.manacero,alexandro.baldassin}@unesp.br

Abstract. *In a technological context, where data are generated exponentially, the financial analysis has gradually become more important to avoid large losses due to fraud. In this paper, we seek to find the segmentation of transactions, through clustering techniques, based on the existence of distinct patterns between legitimate and illegal financial transactions. For this purpose, algorithms were tested and compared in terms of performance, cluster validation, interpretation and understanding, with the last three criteria being used to formulate hypotheses. As a result, a spacial search reduction is expected so that possible frauds can be investigated.*

Resumo. *Em um contexto tecnológico, em que dados são gerados de maneira exponencial, as análises financeiras tem se tornado gradativamente mais importantes para evitar grandes perdas devido às fraudes. Neste trabalho, buscase a segmentação das transações em grupos, por meio de técnicas de agrupamento, com base na existência de padrões distintos entre transações financeiras legítimas e ilegais. Para isto, algoritmos foram testados e comparados em relação ao desempenho, validação do agrupamento, interpretação e compreensão, sendo os três últimos critérios utilizados para a formulação de hipóteses. Como resultado espera-se uma redução do espaço de busca para que possíveis fraudes possam ser investigadas.*

1. Introdução

Considerando-se um contexto tecnológico de geração exponencial de dados, as análises financeiras destacam-se. Isto ocorre uma vez que elas podem evitar grandes prejuízos em instituições deste setor em eventos adversos como as fraudes, reduzindo o impacto, especialmente, após as recentes crises econômicas mundiais.

Analisar dados de clientes comparando-os entre si, buscando agrupar comportamentos similares, permite ao analista identificar riscos, mudanças drásticas ou ocorrências raras. Permite direcionar a análise para grupos específicos que caracterizam maior risco de prejuízo, garantindo maior cobertura e construindo um cenário mais viável, já que as instituições financeiras processam grandes quantidades de transações de clientes em curtos períodos de tempo. Destas transações, um pequeno número é derivado de ações ilegais e, ainda assim, podem causar sérias perdas às empresas [Leite et al. 2017].

Este trabalho busca comparar algoritmos de agrupamento, trazer propostas de interpretabilidade e compreensão de modelos às soluções existentes de detecção de fraudes e avaliar o tempo de execução. Nele, a partir de um conjunto de dados sintéticos, quatro algoritmos são comparados, tendo seu desempenho medido por meio de métricas internas e externas, com relação à sua compreensão, interpretabilidade e tempo de execução. Quanto às métricas internas e externas, estes algoritmos já apresentaram resultados satisfatórios de agrupamento, como será apresentado na Seção 4.

2. Motivação e Objetivos

A segmentação de clientes consiste no processo de dividir os clientes em grupos distintos que compartilham características semelhantes. Como tais características, pode-se citar interesses, comportamento, localização geográfica, entre outros. Com os perfis dos clientes traçados, eles podem auxiliar uma empresa a concentrar recursos em públicos específicos, personalizar campanhas de marketing ou avaliar comportamento de pagamento dos clientes com o intuito de categorizá-los para processamento posterior [Zhou et al. 2021].

Vários trabalhos sobre o assunto podem ser encontrados na literatura em diversos ramos de negócios. Letizio [Letizio 2021] busca garantir que quantias monetárias sejam debitadas e creditadas de maneira correta, protegendo os ativos dos clientes. O autor define as características esperadas para cada cliente e agrupa as semelhantes por meio de algoritmos não supervisionados. Essa abordagem possibilita uma investigação de menor quantidade de registros a fim de rotular transações fraudulentas.

Em outro caso, Nunes, Colliri, Lauretto, Liu e Zhao [Nunes et al. 2021], a partir de uma base de dados abertos de cartões corporativos do governo federal, buscam detectar anomalias e pontos fora da curva (*outliers*) por meio de algoritmos de agrupamento.

Assim, observa-se que a aplicação de métodos com desempenho eficiente para agrupamento de dados em sistemas anti-fraude é um tema bastante relevante. Pontos como aumento nos casos de fraudes, possibilidade de interpretabilidade dos modelos para identificação de fraudes, redução do espaço de busca e possibilidade de comparação de desempenho pelo tempo de execução justificam este trabalho.

3. Metodologia

No trabalho proposto, com o intuito de utilizar a abordagem de agrupamento para redução do espaço de busca, facilitando a análise de possíveis casos de fraudes, foi escolhida uma base de dados do site Kaggle[1]. Essa escolha ocorreu por conta do conjunto ser composto por dados sintéticos (gerados por simulação) referentes à compras digitais (e-commerce) e por sua boa documentação. O conjunto possui 594.643 registros, entre transações normais e fraudulentas, de maneira desbalanceada (menos de 2% dos registros são referentes à fraudes), e um total de 4.112 clientes.

Nesta base foi realizada uma etapa de pré-processamento, melhorando a qualidade dos dados e reduzindo a quantidade de registros por amostragem (foram utilizados cerca de 14.000 registros). Houve a definição de parâmetros ótimos para cada algoritmo de agrupamento (K-means, HDBSCAN, *Agglomerative Clustering* e *Spectral Clustering*).

[1] *"Fraud Detection on Bank Payments"* - `https://www.kaggle.com/turkayavci/fraud-detection-on-bank-payments`

Para a obtenção do número de grupos a serem formados (k), nos algoritmos que requerem tal parâmetro, foi utilizado o método do cotovelo [Kodinariya and Makwana 2013]. No *Spectral Clustering*, para redução dimensional de grafos, foram aplicadas técnicas de *embedding* com o algoritmo de aprendizado profundo *Structural Deep Network Embedding* (SDNE) [Wang et al. 2016], melhorando o desempenho de execução e os resultados obtidos.

Os grupos formados pelos algoritmos tiveram seu desempenho analisado por três métricas internas (métricas que dependem exclusivamente dos grupos formados para seu cálculo): Davies-Bouldin (média de similaridade entre o grupo e seu grupo mais próximo), Silhouette (coeficiente de sobreposição de grupos) e Calinski-Harabasz (medida de densidade e separabilidade entre grupos). Também por duas métricas externas (métricas que dependem de um rótulo para avaliar o agrupamento formado), os índices de Jaccard e Rand Ajustado (ambas métricas de similaridade entre grupos), finalizando aqui as etapas do projeto já realizadas [Zaki et al. 2014]. Todos os códigos foram feitos em Python 3.9.4 (64 bits) e executados em uma máquina com processador Intel Core i7 de 11ª geração, 16GB de memória, SSD de 512GB e placa gráfica Nvidia GeForce MX350.

Como continuidade do projeto, pretende-se realizar a avaliação de cada algoritmo com relação ao agrupamento de fraudes. Serão utilizadas técnicas de interpretabilidade, transformando o contexto não supervisionado, dos algoritmos de agrupamento, em um contexto supervisionado, que possibilite recuperar a importância e o peso atribuídos a cada um dos atributos pelos algoritmos de segmentação [Ismaili et al. 2014] e ferramentas de visualização da informação, com intuito de melhorar a compreensão dos grupos.

As métricas, a interpretabilidade e a visualização da informação permitirão compreender qual algoritmo possui melhor desempenho na detecção de fraudes. Em equilíbrio com o tempo de execução, esta técnica poderá ser expandida para outros contextos.

4. Resultados Preliminares

Os dados já obtidos foram trabalhados em dois cenários, um com balanceamento (equalização da quantidade de transações fraudulentas e não fraudulentas) e outro sem. Para ambos os cenários foram realizadas 50 repetições dos algoritmos. A partir delas, o desempenho foi avaliado com relação à média (μ), ao desvio padrão (σ), à confiança de 95% (IC), calculada pela forma não paramétrica do teste de Bootstrap [Efron 1981], às métricas internas, externas e ao tempo de execução dos algoritmos. Os resultados obtidos são apresentados na Tabela 1. O número ótimo de grupos em todos os algoritmos foi dois.

Tabela 1. Comparativo de métricas internas e externas nos testes.

	Davies Bouldin			**Sillhouette**			**Calinski-Harabasz**			**Jaccard**			**Rand Score Ajustado**			**Tempo**		
Algoritmo	μ	σ	**IC**	μ	σ	**IC**	μ	σ	**IC**	μ	σ	**IC**	μ	σ	**IC**	μ	σ	**IC**
	Balanceado																	
Agglomerative	**0.356**	0.064	[0.25, 0.47]	**0.907**	0.007	[0.89, 0.92]	20.259	2.485	[13812, 23531]	0.274	0.016	[0.23, 0.29]	**0.002**	0.001	[0, 0.0062]	2.487	0.157	[2.25, 2.77]
HDBSCAN	1.081	0.014	[1.05, 1.11]	-0.105	0.010	[-0.12, -0.08]	449	21	[415, 499]	0.160	0.055	[0.11, 0.24]	0.382	0.011	[0.36, 0.40]	1.253	0.076	[1.14, 1.46]
K-means	0.361	0.017	[0.33, 0.39]	**0.907**	0.003	[0.90, 0.91]	**21.786**	1.433	[18690, 24361]	0.276	0.015	[0.23, 0.29]	**0.002**	0.001	[0, 0.004]	**0.061**	0.010	[0.05, 0.08]
Spectral	1.169	0.366	[0.37, 1.91]	0.399	0.088	[0.21, 0.57]	3.660	1.732	[376, 6309]	**0.301**	0.044	[0.18, 0.35]	0.845	0.489	[0, 0.0013]	3.778	1.021	[2.72, 5.40]
	Desbalanceado																	
Agglomerative	**0.268**	0.163	[0.01, 0.57]	**0.971**	0.030	[0.88, 0.99]	9.276	2.920	[5419, 15673]	**0.880**	0.293	[0, 0.98]	0.169	0.176	[0.03, 0.67]	2.499	0.339	[2.09, 3.23]
HDBSCAN	1.387	0.014	[1.37, 1.42]	0.306	0.004	[0.30, 0.31]	1.015	338	[591, 1946]	0.114	0.003	[0.11, 0.12]	0.023	0.002	[0.02, 0.027]	1.077	0.089	[0.96, 1.28]
K-means	0.343	0.126	[0.14, 0.54]	0.968	0.029	[0.90, 0.99]	**9.900**	3.092	[5874, 17714]	**0.880**	0.293	[0, 0.99]	0.200	0.190	[0.05, 0.72]	**0.061**	0.034	[0.04, 0.11]
Spectral	1.122	0.278	[0.46, 1.52]	0.411	0.086	[0.26, 0.55]	4.016	1.892	[751, 8350]	0.511	0.231	[0.10, 0.96]	**0.001**	0.007	[-0.01, 0.02]	3.912	0.917	[2.67, 5.74]

Por meio dos resultados, pode-se concluir que com os dados desbalanceados, os algoritmos K-means e *Agglomerative Clustering* apresentaram melhores métricas (Davies-Bouldin acima de 0, Sillhouete próximos de 1, Calinski-Harabasz elevado, Rand

Ajustado próximo de 0 e Jaccard próximos a 1) do que os algoritmos HDBSCAN e *Spectral Clustering*. O mesmo ocorre para o caso de dados balanceados, porém com queda no índice de Jaccard. Pode-se perceber ainda que o algoritmo de K-means destaca-se, pois além do desempenho de métricas, também tem menor tempo de execução (ao menos 15 vezes menor que os demais), indicando uma boa solução de custo-benefício. Nota-se ainda baixo desvio padrão nas métricas, com exceção de Calinski-Harabasz, e que elas estão entre o intervalo de confiança utilizado, sendo 50 repetições suficientes para avaliar o modelo. Estas métricas serão avaliadas novamente nas demais etapas do projeto.

5. Conclusões

Conclui-se, até o momento, que os objetivos propostos para o trabalho foram cumpridos, como a obtenção do conjunto de dados alvo para testes, a implementação dos métodos de pré-tratamento dos dados e a otimização de parâmetros dos algoritmos.

As próximas etapas estão centralizadas na obtenção de informações relacionadas à interpretabilidade e à explicabilidade dos algoritmos não supervisionados utilizados. O desempenho será validado também, considerando o agrupamento de transações fraudulentas e o tempo de execução dos algoritmos. Ainda como trabalhos futuros, pode-se citar a possibilidade de paralelização e distribuição dos algoritmos que não suportaram a execução com a base de dados completa, verificando qual *speedup* obtido, em cada um, buscando o uso de uma amostra maior de dados ou a utilização de todo o espaço amostral.

Referências

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2):139–158.

Ismaili, O. A., Lemaire, V., and Cornuéjols, A. (2014). A supervised methodology to measure the variables contribution to a clustering. In *International Conference on Neural Information Processing*, pages 159–166. Springer.

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

Leite, R. A., Gschwandtner, T., Miksch, S., Kriglstein, S., Pohl, M., Gstrein, E., and Kuntner, J. (2017). Eva: Visual analytics to identify fraudulent events. *IEEE transactions on visualization and computer graphics*, 24(1):330–339.

Letizio, K. J. (2021). *Combating Financial Fraud: A Machine Learning Approach*. PhD thesis, Utica College.

Nunes, B., Colliri, T., Lauretto, M., Liu, W., and Zhao, L. (2021). Anomaly detection in brazilian federal government purchase cards through unsupervised learning techniques. In *Brazilian Conference on Intelligent Systems*, pages 19–32. Springer.

Wang, D., Cui, P., and Zhu, W. (2016). Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234.

Zaki, M. J., Meira Jr, W., and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Zhou, J., Wei, J., and Xu, B. (2021). Customer segmentation by web content mining. *Journal of Retailing and Consumer Services*, 61:102588.