Uma Análise Comparativa entre Algoritmos de Agrupamentos de Dados

Narciso F. Sousa
Universidade Federal do Rio Grande do Norte (UFRN)
Caicó – RN, Brasil
narcisofariasmg@gmail.com

Huliane M. Silva
Instituto Federal de Educação Ciência e Tecnologia (IFRN)
Caicó – RN, Brasil
huliane.silva@ifrn.edu.br

Arthur C. Gorgônio
Universidade Federal do Rio Grande do Norte (UFRN)
Natal – RN, Brasil
gorgonioarthur@gmail.com

Flavius L. Gorgônio
Universidade Federal do Rio Grande do Norte (UFRN)
Caicó – RN, Brasil
flavius@dct.ufrn.br

***Resumo*—A mineração de dados pode ser definida como um conjunto de técnicas automáticas de exploração de grandes volumes de dados, cujo objetivo é permitir a descoberta de novos padrões e relações que não seriam facilmente detectadas pela visão humana. As diversas ferramentas computacionais de análise e processamento de dados permitem analisar grandes volumes de dados em questões de segundos, porém aplicações reais costumam ser bem mais complexas e possuir bases de dados bem mais desafiadoras do que as comumente apresentadas na literatura. Neste contexto, este trabalho tem como objetivo analisar e comparar algoritmos de agrupamento de dados usando as bases de dados da FCPS (*Fundamental Clustering Problem Suite*) e da ferramenta YADMT (*Yet Another Data Mining Tool*), que simulam diversas situações presentes em problemas reais. No presente estudo, observou-se que os algoritmos hierárquicos alcançaram uma boa eficácia nas métricas de validação de agrupamento de dados utilizadas.**

***Palavras-chave*—agrupamento de dados, análise comparativa, algoritmos de agrupamento**

I. Introdução

Atualmente, as tecnologias computacionais têm proporcionado avanços significativos na sociedade. Estamos diante da era da informação, onde há um crescimento explosivo de dados oriundos de diversas áreas do conhecimento, que são armazenados em grandes repositórios [1]. Neste contexto, surgiu a necessidade de transformar dados em informações úteis, de interesse para empresas, organizações e pesquisadores. Isso, consequentemente, levou ao surgimento de uma área de pesquisa denominada de mineração de dados.

A mineração de dados consiste no processo de extração de informações relevantes em grandes volumes de dados, buscando identificar padrões consistentes que extrapolam a capacidade humana. Esse processo, normalmente, se dá através da aplicação de técnicas de Aprendizado de Máquina como, por exemplo, classificação, agrupamento, regras de associação, sequências temporais, que detectam relacionamentos sistemáticos existentes em bases de dados, auxiliando na descoberta de conhecimento.

As diversas ferramentas computacionais de análise e processamento de dados podem analisar grandes volumes de dados em um curto espaço de tempo. Por outro lado, a utilização destas ferramentas de modo direcionado à obtenção de resultados úteis e práticos, torna-se uma tarefa árdua e desafiadora, que requer especialistas no domínio da aplicação.

Uma das tarefas de mineração de dados é a análise de agrupamentos (do inglês, *clustering analysis*), cujo objetivo é agrupar objetos semelhantes entre si em um determinado grupo, levando em consideração uma ou mais características em comum. Na literatura, é possível encontrar diversos algoritmos de agrupamento com características diferentes. Essas características podem levar a resultados distintos dependendo da aplicação.

Em geral, algoritmos de agrupamento de dados, têm seu desempenho avaliado a partir de bases de dados sintéticas, que nem sempre exploram/expõem os algoritmos em sua totalidade. Diferentemente das sintéticas, as bases de dados reais costumam ser mais complexas e possuir distribuições mais desafiadoras [2], [3].

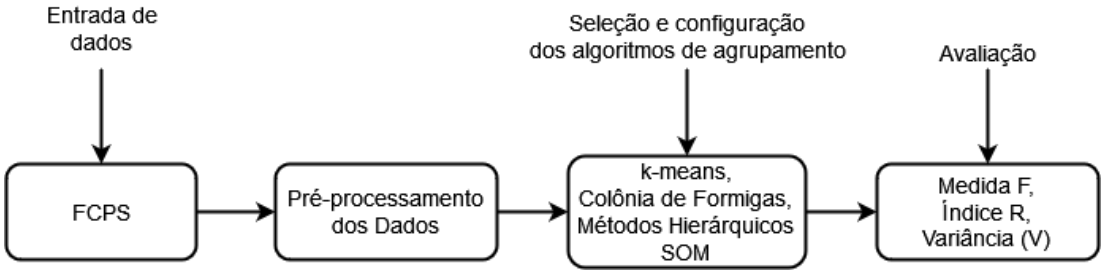
Nesse contexto, o presente trabalho avalia uma ferramenta brasileira de mineração de dados, denominada YADMT (*Yet Another Data Mining Tool*) [4], a partir da utilização de uma *switch* de dados sintéticos que inclui diversas bases de dados desafiadores, descrito como *Fundamental Clustering Problems Suite* (FCPS) [5]. A ideia é aplicar diferentes algoritmos a essas bases de dados, comparar os resultados e verificar a eficácia algoritmos quando aplicadas à bases de dados consideradas desafiadoras.

II. Metodologia Proposta

O presente trabalho descreve uma análise comparativa entre algoritmos de agrupamentos de dados utilizando a Ferramenta YADMT. A Figura 1 ilustra as etapas da metodologia dos experimentos realizados.

Inicialmente, foram selecionadas as bases de dados a serem analisadas. Na etapa seguinte, foi realizado o pré-processamento das bases de dados com o objetivo tornar as bases de dados adequadas à aplicação dos algoritmos. Em seguida, na etapa de mineração de dados, aplicou-se os

Figura 1. Abordagem Proposta



Os Próprios Autores

algoritmos de agrupamento de dados, na tentativa de identificar padrões úteis, difíceis de serem percebidos pela visão humana. Finalmente, analisou-se os agrupamentos gerados por meio de métricas de validação de agrupamentos, com objetivo de verificar a veracidade dos resultados.

III. Configuração Experimental

Neste trabalho, utilizou-se a ferramenta de mineração de dados denominada YADMT, desenvolvida na Universidade Estadual do Oeste do Paraná (Unioeste). Essa ferramenta é livre para uso em ambientes educacionais e possui um módulo de agrupamento de dados, contendo variados algoritmos de agrupamento de dados e alguns índices de validação [6].

A. Bases de Dados

A FCPS é uma coleção de bases de dados tidas como difíceis para testar o desempenho de algoritmos. Os principais desafios que são contemplados nesta *switch*, são: i) falta de separabilidade linear; ii) espaçamento de classe interna diferente ou pequeno; iii) classes definidas pela densidade de dados em vez de espaçamento de dados; iv) nenhuma estrutura de *cluster*; v) *outliers*; e vi) classes que estão em contato [5].

A Tabela I apresenta, sucintamente, as principais características das bases de dados analisadas neste trabalho [7].

B. Algoritmos de Agrupamento de Dados

Os algoritmos utilizados neste trabalho são: colônia de formigas (do inglês, *Ant Colony* – ACO), K-means, métodos hierárquicos e mapas auto-organizáveis (do inglês, *Self-organizing maps* – SOM). Esses algoritmos foram escolhidos por estarem disponíveis na plataforma YADMT e serem amplamente utilizados.

O algoritmo K-means, proposto por [8], é bastante simples e fácil de implementar, consequentemente, bastante conhecido e usado em tarefas de análise de agrupamento. Basicamente, a ideia do K-means resume-se a definição de centróides, que representam a média de um grupo de elementos, e a partir da definição desses centróides serão agrupados N elementos de K grupos, normalmente, baseando-se na distância Euclidiana [9], [10].

Os algoritmos hierárquicos geram uma sequência de partições aninhadas, ou seja, uma estrutura hierárquica do tipo árvore [11]. Basicamente, são divididos em duas categorias: divisivos e aglomerativos. Essas duas categorias diferem na maneira que a estrutura do tipo árvore é construída. Na abordagem divisiva, inicialmente, os objetos estão todos juntos em um único grupo, em seguida é formada uma sequência dividindo os grupos sucessivamente. Na abordagem aglomerativa, cada objeto representa um grupo, em seguida é formada uma sequência de grupos sucessivamente, resultando em um único grupo contendo todos os objetos.

O algoritmo ACO, proposto por [12], é uma metaheurística inspirada no comportamento forrageiro das formigas. Basicamente, o ACO busca encontrar a convergência de uma solução boa o suficiente para o problema, com base em um conjunto de soluções geradas pela comunicação indiretamente entre agentes inteligentes. A versão utilizada neste trabalho contêm algumas modificações propostas por [13].

O algoritmo SOM, proposto por [14], fundamenta-se em um modelo de aprendizagem competitiva. Em geral, ao apresentar uma entrada à rede, os neurônios passam a competir entre si e o vencedor tem seus pesos ajustados, assim como a sua vizinhança. Nisso, o neurônio vencedor e os seus vizinhos tornam-se mais semelhantes ao elemento de entrada e no final do treinamento o mapa fica dividido em regiões, onde cada região possui um conjunto de elementos semelhantes [14].

C. Validação dos Agrupamentos

A validação dos agrupamentos é uma etapa importante no processo de análise de agrupamento, pois é necessário determinar a proximidade existente entre os elementos de um conjunto de dados. Neste trabalho, foram usados a Medida F, o Índice R e a Variância V. A escolha dessas métricas deu-se por avaliarem os agrupamentos de forma diferente, sendo duas internas (Índice R e Variância V) e uma externa (Medida F), e também por já estarem implementadas na ferramenta.

1) A Medida F consiste em realizar uma análise de precisão entre os agrupamentos esperados e os agrupamentos gerados, utilizando conhecimento prévio dos rótulos de uma base de dados [15];
2) O Índice R tem como objetivo calcular a proporção de pares de objetos entre duas partições dos agrupamentos. Ou seja, é uma medida da porcentagem de decisões corretas feitas pelo algoritmo, que concentra-se nas diferenças individuais dos dados de cada agrupamento [16];

Tabela I
BASES DE BASES DE DADOS DA FCPS

Nome	Descrição	Instâncias	Atributos	Classes
Hepta	Claramente definidos e diferentes variações	212	3	7
LSun	Diferentes variações e distâncias	400	2	3
Tetra	Grupos quase se tocando	400	3	4
ChainLink	Linear não separável	1000	3	2
Atom	Variâncias diferentes e linear não separável	800	3	2
Target	Outliers	770	2	6
TwoDiamonds	Bordas dos grupos definidas pela densidade	800	2	2

Fonte: Os Próprios Autores

3) A Variância Intra-Grupos V mensura as diferenças entre os grupos [15].

D. Configuração dos Experimentos

Os experimentos realizados compararam as eficácias dos algoritmos quando aplicados a cada base de dados, através do desempenho obtido pelas métricas: Medida F, Índice R e Variância V. Em cada execução, utilizou-se o número de grupos igual à quantidade de classes conhecida em cada base de dados. Para todos os algoritmos, com exceção do hierárquico aglomerativo (por ser determinístico), foram realizadas 10 execuções, com diferentes inicializações, gerando valores diferentes para as métricas avaliadas. As classes de cada base de dados foram omitidas no processo de agrupamento de dados.

Em cada execução, para definir o número de neurônios no mapa no algoritmo SOM, utilizou-se a quantidade de instâncias das bases de dados. Nos demais parâmetros deste algoritmo utilizou-se as configurações *default*, de acordo com a ferramenta YADMT. No algoritmo ACO, a cada execução, os dados são colocados aleatoriamente na grade. Para o algoritmo K-means, a cada execução, os centróides são inicializados aleatoriamente.

Para as comparações foram utilizadas diferentes versões dos algoritmos do SOM, ACO, MH. Os algoritmos a partir do SOM foram: 1D-SOM, Matriz e Densidade, Metodologia de Vesanto e Alhoniemi, SL-SOM e Componentes Conectados. Os algoritmos de Métodos Hierárquicos utilizaram a recuperação de agrupamento tais como: Ligação Simples, Ligação Completa, Ligação Média e Método Ward. bem como, o algoritmo colônia de formigas. Em todos algoritmos foram selecionadas as melhores versões para realizar as comparações.

IV. ANÁLISE DOS RESULTADOS

A Tabela II apresenta os resultados dos testes comparativos executados a partir dos algoritmos selecionados para este trabalho. O critério usado na análise comparativa foi a eficácia de cada algoritmo, levando em consideração seu desempenho obtido através dos índices de validação Medida F e Índice R. Os valores mais altos desses índices indicam alto grau de similaridade entre os dados nos grupos. Valores mais próximos a 1,0 indicam melhor agrupamento, pois seus valores variam no intervalo [0,1]. Além disso, utilizou-se ainda a variância V, que demonstra o afastamento da média dos dados do conjunto analisado, por isso quanto menor, melhor é o agrupamento.

Tabela II
RESULTADOS

Base de Dados	Métricas	ACO	SOM	MH	K-means
Atom	Índice R (↑)	0,626	0,818	1,000	0,597
	Medida F (↑)	0,530	0,818	1,000	0,697
	Variância V (↓)	48342,1	35044,9	13333,2	41721,3
Chainlink	Índice R (↑)	0,685	0,713	1,000	0,543
	Medida F (↑)	0,572	0,669	1,000	0,648
	Variância V (↓)	74850,9	53740,1	20833	79002,8
Hepta	Índice R (↑)	0,347	0,940	1,000	1,000
	Medida F (↑)	0,776	0,973	1,000	1,000
	Variância V (↓)	3374,07	1005,75	76,4764	76,4764
Lsun	Índice R (↑)	0,517	0,912	1,000	0,746
	Medida F (↑)	0,469	0,929	1,000	0,723
	Variância V (↓)	13017,4	2701,42	2083,25	10168,8
Target	Índice R (↑)	0,650	0,613	1,000	0,701
	Medida F (↑)	0,487	0,553	1,000	0,820
	Variância V (↓)	49186,7	41870	13371,3	13826,4
Tetra	Índice R (↑)	0,448	0,619	1,000	1,000
	Medida F (↑)	0,653	0,712	1,000	1,000
	Variância V (↓)	10711	10568,3	833,25	833,25
TwoDiamonds	Índice R (↑)	0,714	0,704	0,745	1,000
	Medida F (↑)	0,593	0,601	0,634	1,000
	Variância V (↓)	45297,4	40750,6	39583,6	13333,2
Total de Vitórias		0	0	18	9

Pode-se observar que os algoritmos hierárquicos obtiveram performance igual ou superior em seis das sete bases de dados analisadas, seguidos pelo algoritmo K-means, que obteve desempenho igual ou superior em três dos sete conjuntos de dados analisados. Os algoritmos SOM e ACO, apesar de conseguirem identificar os agrupamentos, não conseguiram se destacar positivamente em nenhum dos casos.

Uma particularidade importante que merece ser destacada é que os algoritmos hierárquicos possuem uma vantagem sobre os demais algoritmos analisados que é inerente ao seu funcionamento. Por não necessitarem de um valor que indique o número de agrupamentos esperado, é sempre possível escolher posteriormente um número de agrupamentos que otimize a escolha, isso lhes garante uma vantagem adicional. Entretanto, a sua complexidade (quadrática) é alta, por necessitar comparar as distâncias de todos os elementos entre si, o que pode torná-lo inviável para utilização em bases de dados de alta dimensionalidade.

O K-means, por sua vez, demonstrou ser uma escolha intermediária, tendo desempenho próximo ao obtido pelos algoritmos hierárquicos, mas com uma complexidade computacional menor (linear), o que lhe permite a utilização com volumes de dados de grande dimensionalidade. Entretanto, a literatura alerta para sua sensibilidade relacionada à inicialização dos

centróides. Outra limitação é a necessidade de se conhecer, antecipadamente, a quantidade de agrupamentos existentes nos dados, o que nem sempre é uma informação disponível em aplicações do mundo real [17].

Os mapas auto organizáveis (SOM) também obtiveram valores próximos aos obtidos pelos métodos hierárquicos e pelo K-means em alguns conjuntos de dados, mostrando-se como uma opção competitiva e com complexidade de ordem linear. Entretanto, o seu resultado final é uma matriz de distâncias e há a necessidade de se executar uma etapa de pós-processamento para identificação dos agrupamentos. Por isso, tanto o algoritmo SOM, quanto o ACO apresentam como vantagem a possibilidade de se executar tais métodos com diferentes valores de k e utilizar métricas (índices) de validação de agrupamentos para a escolha da melhor aproximação.

É importante destacar ainda que este trabalho limitou-se em avaliar a eficácia dos algoritmos citados, levando-se em consideração apenas o desempenho obtido pelos mesmos através dos índices de validação. Os resultados aqui apresentados utilizaram apenas o conjunto de bases de dados da FCPS, sendo assim, tais resultados podem não ser os mesmos quando utilizadas outras bases de dados.

V. Considerações Finais e Trabalhos Futuros

Este artigo propôs uma análise comparativa entre alguns algoritmos de agrupamento da ferramenta YADMT, usando conjuntos de dados considerados desafiadores (*Fundamental Clustering and Projection Suite*). Neste estudo, observou-se que os algoritmos hierárquicos alcançaram uma boa eficácia em relação aos demais, na tarefa de agrupamento de dados. É importante destacar que os algoritmos hierárquicos são métodos simples que não requerem informações *a priori* a respeito do problema, como por exemplo o número de grupos, que pode ser inferido através da análise do dendograma. Entretanto, os demais algoritmos de agrupamento utilizados neste trabalho, que requerem o número de grupos *a priori*.

Os resultados experimentais descritos podem ser usados para guiar a escolha de algoritmos de agrupamento de dados em aplicações do mundo real. Todavia, esses resultados não podem ser generalizados para quaisquer outras bases de dados, haja vista que tais escolhas dependem de alguns critérios, tais como: o analista do problema, condições técnicas, estatísticas, configurações e parâmetros usados em todo processo de análise de agrupamento.

Como trabalhos futuros, pode-se sugerir a alteração das configurações dos algoritmos (ACO, SOM, MH e K-means) utilizados, uma vez que quaisquer alterações nos parâmetros resultam em novos resultados. Além disso, podem ser estudados outros algoritmos de agrupamento, tais como: DBSCAN e *Expectation-Maximization* (EM). Também podem ser avaliados outras métricas de validação de agrupamentos para a escolha da melhor aproximação (*Silhouette*, *Davies–Bouldin* e *Adjust Rand*). Por fim, pode-se comparar as implementações dos mesmos algoritmos de agrupamento que estejam disponíveis em outras ferramentas de mineração de dados.

Referências

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques, Third Edition*. Waltham, Mass: Morgan Kaufmann Publishers, 2012, vol. 3rd.

[2] R. Narayanan, B. Ozisikyilmaz, J. Zambreno, G. Memik, and A. Choudhary, "Minebench: A benchmark suite for data mining workloads," in *2006 IEEE International Symposium on Workload Characterization*, 2006, pp. 182–188.

[3] W. Konen, P. Koch, O. Flasch, T. Bartz-Beielstein, M. Friese, and B. Naujoks, "Tuned data mining: a benchmark study on different tuners," in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, 2011, pp. 1995–2002.

[4] E. Benfatti, F. Bonifacio, A. Girardello, and C. Boscarioli, "Descrição da arquitetura e projeto da ferramenta yadmt-yet another data mining tool," *Relatório Técnico*, no. 01, 2011.

[5] A. Ultsch and J. Lötsch, "The fundamental clustering and projection suite (fcps): A dataset collection to test the performance of clustering and data projection algorithms," *Data*, vol. 5, no. 1, 2020. [Online]. Available: https://www.mdpi.com/2306-5729/5/1/13

[6] C. Boscarioli, M. F. Teixeira, R. Villwock, and T. M. Faino, "O módulo de agrupamento de dados da ferramenta yadmt," *V Epac - Encontro Paranaense de Computação*, 2013.

[7] A. Ultsch, "Clustering wih som," *In Proc. Workshop on Self-Organizing Maps*, 2005.

[8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.

[9] Y. Li and H. Wu, "A clustering method based on k-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 12 2012.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[11] A. C. Lorena, J. Gama, and K. Faceli, *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC, 2000.

[12] J.-L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chrétien, "The dynamics of collective sorting robot-like ants and ant-like robots," in *Proceedings of the International Conference on Simulation of Adaptive Behavior*, 01 1990, pp. 356–363.

[13] R. Villwock, "Técnivas de agrupamento e de hierarquização no contexto de kdd - aplicação a dados temporais de instrumentação geotécnica-estrututal da usina hidrelétrica de itaipu," *Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração de Programação Matemática, dos Setores de Tecnologia e de Ciências Exatas, da Universidade Federal do Paraná*, 2009.

[14] T. Kohonen and T. Honkela, "Kohonen network," *Scholarpedia*, vol. 2, no. 1, p. 1568, 2007, revision #127841.

[15] A. A. Knob, "Formas de mapeamento do problema cash para agrupamento de dados," *Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná*, 2015.

[16] V. A. Padilha and A. C. P. L. F. Carvalho, "Mineração de dados em python," in *Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo*, 2017.

[17] N. F. de Sousa, F. Gorgônio, and H. Medeiros, "Um estudo comparativo entre algoritmos de agrupamentos de dados usando a ferramenta yadmt," *Escola Regional do Ceará, Maranhão e Piauí*, 2021.