Um Estudo Comparativo entre Algoritmos de Agrupamentos de Dados Usando a Ferramenta YADMT

Narciso F. Sousa[1], **Flavius L. Gorgônio**[1], **Huliane M. Silva**[2]

[1]Universidade Federal do Rio Grande do Norte (UFRN)
Rua Joaquim Gregório, 296 - Penedo, CEP 59300-000 – Caicó – RN – Brasil

[2]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN)
RN 288, S/N - Nova Caicó, CEP 59300-000 – Caicó – RN – Brasil

narcisofariasmg@gmail.com, flavius@dct.ufrn.br, huliane.silva@ifrn.edu.br

Abstract. *The need to transform data into information and information into knowledge led to the emergence of the data mining area, whose objective is to provide techniques for interpreting large volumes of data. Although current computational tools for analyzing and processing information can analyze huge volumes of data in a matter of seconds, real-world applications tend to be much more complex and have much more challenging databases than those commonly presented in the literature. This work presents a comparative study between data clustering algorithms from Fundamental Clustering Problem Suite (FCPS) databases and the Yet Another Data Mining Tool (YADMT), which simulate various situations present in real world problems. The algorithms chosen in this research were: ant colony, k-means, self-organizing maps and hierarchical methods. For their evaluation, the F-Measure, the R-Index and the Intra-groups Variance were used.*

Resumo. *A necessidade de transformar dados em informação e informação em conhecimento, impulsionou o surgimento da área de mineração de dados, cujo objetivo é fornecer técnicas de interpretação de grandes volumes de dados. Embora as atuais ferramentas computacionais de análise e processamento de informação possam analisar imensos volumes de dados em questões de segundos, aplicações do mundo real costumam ser bem mais complexas e possuir bases de dados muito mais desafiadoras do que as comumente apresentadas na literatura. Este trabalho apresenta um estudo comparativo entre algoritmos de agrupamento de dados a partir de bases de dados da FCPS (Fundamental Clustering Problem Suite) e da ferramenta YADMT (Yet Another Data Mining Tool), que simulam variadas situações presentes em problemas do mundo real. Os algoritmos escolhidos nesta pesquisa foram: colônia de formigas, k-means, mapas auto-organizáveis e métodos hierárquicos. Para avaliação dos mesmos foram utilizados a Medida F, o Índice R e a Variância Intra-Grupos.*

1. Introdução

Analisando-se a sociedade contemporânea, é inegável que a utilização de tecnologias de informação e comunicação têm proporcionado grandes avanços produtivos e significativos para toda a humanidade. De fato, vive-se atualmente a era da informação, onde o crescimento exponencial de dados, acumulados em dispositivos de armazenamento de baixo

custo e compartilhados através de redes de computadores, tornam a web e as técnicas de análise de dados em ferramentas protagonistas da gestão de informação na computação moderna [Han et al. 2012].

A necessidade de explorar imensas bases dados em busca de conectar informações e obter conhecimento, levou ao surgimento da área de mineração de dados no início da década de 1990. A mineração de dados contribui para suprir a carência de técnicas de interpretação de grandes volumes de dados, através de sistemas computacionais que buscam soluções para problemas que ultrapassam a capacidade humana de resolvê-los [Fayyad et al. 1996].

Muito embora as atuais ferramentas computacionais de análise e processamento de informação possam analisar imensos volumes de dados em questões de segundos, o uso dessas ferramentas de modo objetivo e direcionado à obtenção de resultados úteis e práticos ainda é um desafio que requer profissionais devidamente capacitados e qualificados, com conhecimento especializado sobre o modelo de negócios a ser analisado. Um algoritmo (ou ferramenta) que obtém bons resultados em uma determinada aplicação, pode não ser tão eficaz ou eficiente em outro contexto e a escolha da ferramenta mais adequada nem sempre é uma tarefa trivial.

Além disso, aplicações do mundo real costumam ser bem mais complexas e possuir bases de dados muito mais desafiadoras do que as comumente apresentadas na literatura [Narayanan et al. 2006, Konen et al. 2011]. Métodos, técnicas e algoritmos de análise de dados, na maior parte das vezes, têm seu desempenho avaliado a partir de conjuntos de dados de teste (*benchmarks*), que nem sempre incluem situações situações limite encontradas em bases de dados reais, o que dificulta ainda mais a escolha das ferramentas mais adequadas em um processo de mineração de dados.

Este trabalho tem como objetivo geral realizar uma análise comparativa da eficácia de alguns algoritmos de agrupamentos de dados, a partir de um *benchmark* que inclui diversos conjuntos de dados que simulam situações-limite, descrito como *Fundamental Clustering Problems Suite* (FCPS) [Ultsch and Lötsch 2020]. A proposta é comparar diferentes algoritmos de mineração de dados a partir desse *benchmark*, avaliar os resultados e verificar a eficácia desses algoritmos quando aplicados a bases de dados consideradas desafiadores.

2. Fundamentação Teórica

2.1. Análise de Agrupamentos

A tarefa de análise de agrupamento (algumas vezes denominada *clustering*), técnica de interesse desta pesquisa, é uma estratégia computacional de identificação de objetos (dados) em agrupamentos (grupos), na qual cada agrupamento deverá conter objetos com a máxima semelhança entre si e consideralmente diferentes daqueles associados a outros agrupamentos. Para se mensurar a semelhança entre objetos, faz-se o uso de alguma métrica que simule uma função de distância estatística.

O cálculo da distância entre os atributos de dois objetos é uma métrica comum para avaliar o agrupamento de dados numéricos. Em geral, quando o valor de tal distância é pequeno, indica maior semelhança entre os dados, sendo a recíproca verdadeira. Existem

várias métricas de distância comumente sugeridas na literatura, por exemplo: Euclidiana, Manhattan, Minkowski, Correlação, Mahalanobis, entre outras.

Dentre as métricas apresentadas, a distância Euclidiana é a mais comum, sendo a mais utilizada nos mais diversos algoritmos de agrupamentos de dados. Considerando-se um plano euclidiano bidimensional, onde $P = (p_x, p_y)$ e $Q = (q_x, q_y)$ são pontos arbitrariamente dispostos nesse plano, x e y são coordenadas cartesianas no plano. A distância Euclidiana pode ser computada como sendo:

$$d_{PQ} = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \tag{1}$$

A Eq. (1) pode ainda ser generalizada como uma medida de distância no plano n-dimensional, incluindo-se as componentes dos outros planos. Este trabalho utilizou a medida de distância Euclidiana como métrica de similaridade para os algoritmos K-means, Métodos Hierárquicos, Mapas Auto-Organizáveis e Colônia de Formigas.

2.2. Algoritmos de Agrupamento de Dados

A análise de agrupamentos compõe um conjunto de técnicas estatísticas multivariadas que possibilitam analisar um conjunto de objetos considerando, simultaneamente, múltiplas características de cada um deles, tendo como objetivo identificar subconjuntos significativos de itens, mutuamente excludentes, com base nas similaridades existentes entre as características de cada objeto [Hair et al. 2009].

O K-means é um algoritmo de agrupamento não-hierárquico, proposto por J. B. MacQueen a partir de técnicas de quantização vetorial, sendo talvez o mais popular na área de mineração de dados e reconhecimento de padrões, devido a sua rapidez e facilidade de implementação [MacQueen et al. 1967].

Os algoritmos baseados em Métodos Hierárquicos (MH) descobrem agrupamentos e um possível número de grupos utilizando técnicas aglomerativas (que usam uma série de fusões dos dados sucessivamente) ou técnicas divisivas (que usam uma série sucessivamente de divisão dos dados [Linden 2009]. Estes métodos hierárquicos de agrupamentos, diferem-se entre si pela maneira de medir a distância entre dois agrupamentos, alguns são descritas como: Ligação Simples, Ligação Média, Ligação Completa e Ward.

O algoritmo de Colônia de Formigas (ACO) é baseado na maneira como as formigas da espécie *Pheidole pallidula* organizam seus corpos, quando mortos, em grupos de itens similares, além de observações sobre a forma como elas agrupam seus ovos (larvas) de acordo seu tamanho. O algoritmo criado a partir deste modelo, foi introduzido por [Deneubourg et al. 1990], para ser focado em agrupar objetos usando um grupo de robôs do mundo real. A implementação do ACO utilizada nesta pesquisa contém algumas modificações propostas por [Villwock 2009].

Os Mapas Auto-Organizáveis (*Self-Organizing Map* - SOM) foram propostos por T. Kohonen, inspirado no funcionamento dos neurônios relacionados ao córtex cerebral humano. Esta abstração possui uma topologia com duas camadas, uma de entrada e outra de saída, sendo que há uma ligação sináptica das unidades de entrada com todas as de saída. Os agrupamentos acontecem da forma em que os elementos com características parecidas são projetados próximo uns aos outros na camada de saída [Kohonen and Honkela 2007].

2.3. Trabalhos Relacionados

Entre trabalhos relacionados a esta pesquisa, [Teixeira 2013] apresentou um comparação entre os algoritmos K-means, Colônia de Formigas os Métodos Hierárquicos. Para todos os algoritmos foram utilizadas seis bases de dados e nenhum dos métodos apresentou resultados satisfatórios levando em consideração a qualidade dos agrupamentos, com exceção do K-means, para uma determinada base de dados.

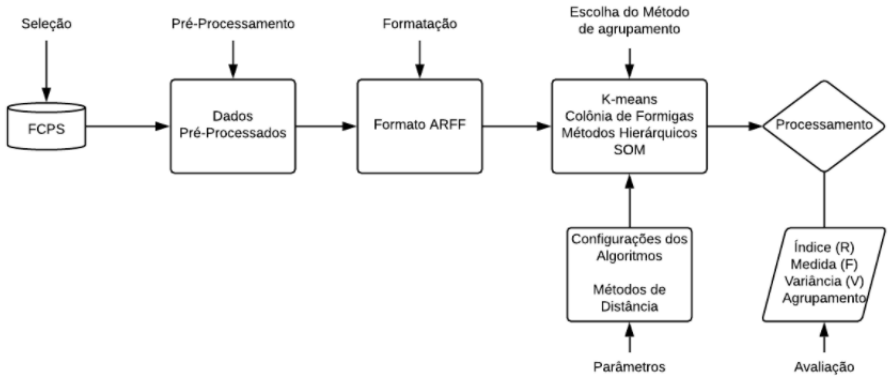
Em outra abordagem comparativa, [Faino 2013], fez comparações dos algoritmos de agrupamento da ferramenta YADMT, entre eles: mapas SOM, Colônia de Formigas e K-means, utilizando bases de dados reais. Das cinco bases de dados experimentadas, o agrupamento a partir do SOM obteve melhor resultado em quatro. Neste trabalho, os autores sugerem que trabalhos futuros investiguem os algoritmos utilizando outras bases de dados. Diferente dos trabalhos anteriores, que utilizam bases de dados comumente referenciadas na literatura, esta pesquisa objetiva comparar os algoritmos K-means, MH, ACO e SOM a partir de bases de dados que simulam problemas de difícil agrupamento, como as presentes no FCPS.

3. Materiais e Métodos

3.1. Abordagem Metodológica

Conforme já mencionado, este trabalho apresenta um estudo comparativo entre algoritmos de agrupamentos de dados utilizando a ferramenta YADMT. Tal ferramenta de mineração de dados, desenvolvida na Universidade Estadual do Oeste do Paraná (Unioeste), é um software 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 [Boscarioli et al. 2013]. A Figura 1 ilustra as etapas da abordagem proposta neste trabalho.

Figura 1. Abordagem Proposta



Adaptado

Inicialmente, escolhe-se a base a ser analisada. Em seguida, realiza-se o pré-processamento da base de dados com o objetivo de eliminar ruídos, dados faltosos e valores ilegítimos, ou seja, tornar a base de dados adequada para aplicação dos algoritmos. Após a etapa de pré-processamento, realiza-se a formatação da base de dados de acordo com o padrão aceitável pela ferramenta. Na etapa de mineração de dados aplica-se algoritmos de aprendizado de máquina, na tentativa de extrair padrões imperceptível pela

visão humana. Por último, na etapa de pós-processamento, avalia-se o conhecimento descoberto, verificando se os resultados são satisfatórios.

3.2. Bases de Dados

Na análise experimental realizada neste trabalho, utilizou-se um conjunto de base de dados da FCPS, descritos como problemas desafiadores. De acordo com [Ultsch and Lötsch 2020], a FCPS é uma coleção de bases de dados para testar o desempenho de algoritmos, que aborda diversos desafios para os algoritmos, como: falta de separabilidade linear, espaçamento de classe interna diferente ou pequeno, classes definidas pela densidade de dados em vez de espaçamento de dados, nenhuma estrutura de *cluster*, *outliers* ou classes que estão em contato. Para comparar os algoritmos de agrupamento foram utilizadas sete bases de dados. A Tabela 1 descreve de forma sucinta as principais características conjunto de bases de dados com seus respectivos nomes, descrições, instâncias (INS), atributos (ATR) e Classes [Ultsch and Lötsch 2020].

Tabela 1. Conjunto de Bases de Dados da FCPS

Nome	Descrição	INS	ATR	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

3.3. Validação dos Agrupamentos

A validação dos agrupamento 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 Medida *F* realiza uma análise de precisão entre os agrupamentos desejados e os agrupamentos gerados, utilizando conhecimento prévio acerca dos rótulos de uma base de dados[Knob 2015]. Para o cálculo da medida *F* utiliza os itens a seguir:

- *r*: memória
- *p*: precisão
- *i*: contém *n*i agrupamentos esperados
- *j*: contém *nj* agrupamentos do algoritmo
- *nij*: é o número de agrupamentos da classe *i* que estão no grupo *j*
- *r(i,j)* ou $\frac{nij}{ni}$: é a memória *r* para cada classe *i*
- *p(i,j)* ou $\frac{nij}{nj}$: é a precisão *p*, para cada grupo *j*

calcula-se primeiro a *F*:

$$F(i,j) = \frac{(b^2+1).p(i,j).r(i,j)}{b^2.p(i,j) + r(i,j)}, \tag{2}$$

para manter os pesos iguais para *p* e *r* o valor de *b* deve ser "1", seguindo a seguir a equação da medida *F* [Handl et al. 2003]:

$$F = \sum i \frac{ni}{n} max_j[F(i,j)]. \tag{3}$$

O índice *R*, é uma medida da porcentagem de decisões corretas feitas pelo algoritmo e concentra-se em diferenças individuais dos dados em cada grupo [Padilha and Carvalho 2017]. Os critérios de validação externa em várias aplicações práticas usa uma partição da solução obtida do algoritmo, equacionado da seguinte forma:

$$C^{ob} = {C_1}^{ob}, ..., {C_k}^{ob}, \tag{4}$$

por fim a partição restante representa a solução a ser comparada, ou seja, a solução do conjunto de dados estudado, definido por:

$$C^{ref} = {C_1}^{ref}, ..., {C_q}^{ref}. \tag{5}$$

As variáveis que definem os critério de semelhança são:

- a_{11}: quantidade de pares de objetos em um mesmo grupo em C^{ob} e C^{ref};
- a_{10}: quantidade de pares de objetos em um mesmo grupo em C^{ob} mas em um mesmo grupo em C^{ref};
- a_{00}: quantidade de pares de objetos pertecentes a grupos diferentes tanto em C^{ob} quanto em C^{ref}.

O índice *R* calcula a proporção de acordos no agrupamento de pares de objetos entre duas partições (a_{11} e a_{00}) em relação ao total de pares possíveis de objetos, ou seja [Padilha and Carvalho 2017]:

$$R(C^{ob}, C^{ref}) = \frac{a_{11} + a_{00}}{a_{11} + a_{10} + a_{01} + a_{11}} = \frac{a_{11} + a_{00}}{\binom{n}{2}}. \tag{6}$$

A Variância Intra-Grupos *V*, concentra-se nas diferenças entre os grupos [Knob 2015]. A equação da Variância é demonstrada na equação 7.

$$V = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2, \tag{7}$$

onde *k* é o número de grupos, *x* é o elemento pertencente ao agrupamento *Ci* e m_i é o centróide do grupo *i*.

3.4. Configuração dos Experimentos

Os experimentos realizados compararam a performance de cada algoritmo para cada base de dados, através do desempenho obtido pela Medida *F*, Índice *R* e Variância *V*. Nesta etapa, cada algoritmo foi executado em cada base de dados, usando o número de grupos igual à quantidade de classes conhecida. Para os algoritmos não determinístico, foram realizadas 10 execuções, com diferentes inicializações, computando-se a média. As classes de cada conjunto de dados foram omitidas no processo de agrupamento.

Para o algoritmo SOM, foram utilizadas a quantidade de instâncias das bases de dados para definir o número de neurônios no mapa utilizado. As configurações deste algoritmo foram deixadas da forma padrão, ou seja, não houve mudanças nas configurações *default* contidas na ferramenta de mineração de dados.

4. Resultados

A Tabela 2 apresenta os resultados da performance dos algoritmos selecionados, a partir da sua eficácia, levando em consideração seu desempenho obtido através dos índices de validação. Os valores mais altos dos índices de avaliação medida *F* e medida *R* 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]. E a variância *V* demonstra o afastamento da média dos dados do conjunto analisado, por isso quanto menor, melhor é o agrupamento.

Tabela 2. Resultados

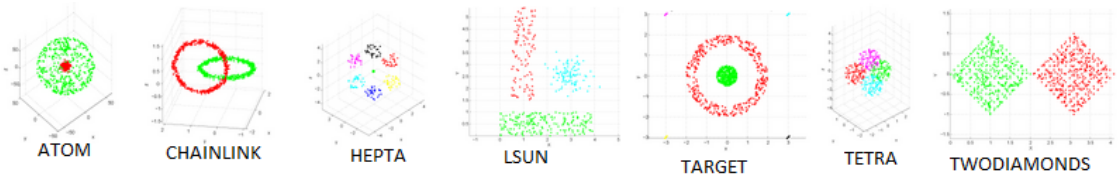
Base de Dados	Algoritmos	Índice R	Medida F	Variância V
Atom	ACO	0,626	0,530	48342.1
	SOM	0,818	0,818	35044.9
	MH	1,0	1,0	13333.2
	K-means	0,597	0,697	41721.3
Chainlink	ACO	0,685	0,572	74850,9
	SOM	0,713	0,669	53740,1
	MH	1,0	1,0	20833,0
	K-means	0,543	0,648	79002,8
Hepta	ACO	0,347	0,776	3374,07
	SOM	0,94	0,973	1005,75
	MH	1,0	1,0	76,4764
	K-means	1,0	1,0	76,4764
Lsun	ACO	0,517	0,469	13017,4
	SOM	0,912	0,929	2701,42
	MH	1,0	1,0	2083,25
	K-means	0,746	0,723	10168,8
Target	ACO	0,65	0,487	49186,7
	SOM	0,613	0,553	41870,0
	MH	1,0	1,0	13371,3
	K-means	0,701	0,82	13826,4
Tetra	ACO	0,448	0,653	10711,0
	SOM	0,619	0,712	10568,3
	MH	1,0	1,0	833,250
	K-means	1,0	1,0	833,200
TwoDiamonds	ACO	0,714	0,593	45297,4
	SOM	0,704	0,601	40750,6
	MH	0,745	0,634	39583,6
	K-means	1,0	1,0	13333,2

Pode-se observar que os algoritmos hierárquicos identificaram seis agrupamentos corretamente, seguido do algoritmo K-means que identificou três agrupamentos, de acordo com a quantidade de classes conhecidas *a priori*, enquanto que os algoritmos SOM e ACO não conseguiram identificar os agrupamentos corretamente.

Portanto, a abordagem proposta neste trabalho identificou que os algoritmos hierárquicos alcançaram uma boa eficácia, exceto em razão das dificuldades encontradas em relação à base de dados TwoDiamonds, onde o fator determinante para distinguir um agrupamento do outro é a borda do cluster definida pela densidade. O algoritmo K-means apresentou dificuldades para identificar corretamente as bases de dados onde há

clusters não esféricos (Atom, Chainlink, Lsun e Target), sendo um ponto crítico a considerar quando atentamos para os resultados da Tabela 2 e visualização das bases de dados na Figura 2.

Figura 2. Visualização das Bases de Dados [Ultsch 2005]



Entretanto, uma limitação inerente à utilização dos métodos hierárquicos, assim como do algoritmo K-means, é a necessidade de se conhecer, antecipadamente, a quantidade de agrupamentos existente nos dados, o que nem sempre é uma informação disponível em aplicações do mundo real. Nesses casos, os algoritmos SOM e 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.

Os algoritmos SOM e ACO não conseguiram identificar os agrupamentos com a mesma precisão dos demais, porém ainda assim obtiveram resultados consideráveis, sendo o SOM mais eficaz que ACO. Situações levantadas nesta pesquisa apontaram que, as performances destes algoritmos está paralelamente associado a complexidade de definir suas melhores configurações e parâmetros, o que infere diretamente em seus resultados.

Sobre a ferramenta YADMT, vale ressaltar que a mesma possui usabilidade bastante intuitiva, além de já possuir os índices de validação incluídos, o que torna mais rápida a análise e comparação dos resultados obtidos. A principal contribuição deste trabalho foi apresentar resultados comparativos que permitem guiar a escolha de algoritmos de análise de agrupamento em aplicações do mundo real. Além disso, mostra que os algoritmos hierárquicos tendem a apresentar melhor performance, quando utilizada a medida de distância euclidiana em bases de dados com baixa dimensionalidade, enquanto que o K-means enfrenta dificuldades com bases cujos agrupamentos não são esféricos.

Como trabalho futuros, pode-se utilizar os resultados deste experimento como referência para futuras comparações, alterar as configurações do algoritmos utilizados, testar outros algoritmos de agrupamento e outras bases de dados. Além disso, pode-se comparar implementações disponíveis em outras ferramentas de mineração de dados mantendo como base de estudo, o conjunto de bases da FCPS, descritas na literatura como desafiadoras.

Referências

Boscarioli, C., Teixeira, M. F., Villwock, R., and Faino, T. M. (2013). O módulo de agrupamento de dados da ferramenta yadmt. *V EPAC Enc. Paranaense de Computação*.

Deneubourg, J.-L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., and Chrétien, L. (1990). The dynamics of collective sorting robot-like ants and ant-like robots. In *Proc of the Int Conference on Simulation of Adaptive Behavior*, pages 356–363.

Faino, T. M. (2013). Agrupamento de dados a partir de mapas auto-organizáveis na ferramenta yadmt. *CCET, UNIOESTE, PR*.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman editora.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques, Third Edition*, volume 3rd. Morgan Kaufmann Publishers, Waltham, Mass.

Handl, J., Knowles, J., and Dorigo, M. (2003). On the performance of ant-based clustering. In *On the Performance of Ant-based Clustering*, volume 104, pages 204–213.

Knob, A. A. (2015). Formas de mapeamento do problema cash para agrupamento de dados. *CCET, UNIOESTE, PR*.

Kohonen, T. and Honkela, T. (2007). Kohonen network. *Scholarpedia*, 2(1):1568. revision #127841.

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

Linden, R. (2009). Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, 1.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Narayanan, R., Ozisikyilmaz, B., Zambreno, J., Memik, G., and Choudhary, A. (2006). Minebench: A benchmark suite for data mining workloads. In *2006 IEEE International Symposium on Workload Characterization*, pages 182–188.

Padilha, V. A. and Carvalho, A. C. P. L. F. (2017). Mineração de dados em python. In *Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo*.

Teixeira, M. F. (2013). Agrupamento e visualização de dados: Estudo e implementações para a ferramenta yadmt. *Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná*.

Ultsch, A. (2005). Clustering wih som. *In Proc. Workshop on Self-Organizing Maps*.

Ultsch, A. and Lötsch, J. (2020). The fundamental clustering and projection suite (fcps): A dataset collection to test the performance of clustering and data projection algorithms. *Data*, 5(1).

Villwock, R. (2009). Técnivas de agrupamento e de hierarquização no contexto de kdd - aplicação a dados temporais de instrumentação geotécnica-estrututal da usina hidrelétrica de itaipu. *PPGMNE, UFPR*.