Universidade Estadual de Campinas

Departamento de Estatística

ME731 - Métodos em Análise Multivariada

Professor Aluísio de Souza Pinheiro

Definição, Aplicação e Características do Particionamento em Medoides (PAM)

Lucas Perondi Kist - 236202

Campinas - SP

2023

1 Introdução

A análise de dados multivariados necessita, em vários momentos, de técnicas que possibilitem o agrupamento de observações semelhantes. Para isso, são utilizados alguns métodos, como K-médias, PAM, CLARA, dentre outros, que realizam essa separação em categorias. Nesse sentido, será definido e discutido o funcionamento do Particionamento em Medoides (PAM), também conhecido como K-medoides, além de ilustrar seu funcionamento em um conjunto de dados.

A aplicação será realizada no conjunto de dados disponibilizado por Wang (2020), que contém características a respeito de 178 vinhos produzidos na Itália. O objetivo é agrupá-los segundo atributos semelhantes, de modo a auxiliar a criação de um certo número de linhas de vinho, o qual também deve ser determinado.

2 Metodologia

2.1 Particionamento em Medoides (PAM)

2.1.1 Definição

O Particionamento em Medoides (do inglês, Partitioning Around Medoids - PAM) é uma técnica de agrupamento de dados contínuos multivariados que objetiva agregá-los em um determinado número de grupos. Para isso, são utilizadas observações do conjunto de dados como centros de cada $cluster$ (medoides), de forma que seja minimizada a soma das "distâncias" das observações a eles.

Formalmente, segundo Kaufman e Rousseeuw (2009), suponha que exista uma matriz de dados, denotada por $\boldsymbol{X}$, da forma:

$$\boldsymbol{X}_{p\times n} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$$

onde

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T,\ i = 1, 2, ..., n.$$

Adicionalmente, o objetivo é agrupar os dados em $k \geq 2$ grupos de acordo com as p medidas quantitativas de cada observação.

Além disso, seja $d(i, j)$ a dissimilaridade entre os objetos $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$, $i, j = 1, 2, ..., n$. Define-se y_i como a variável indicadora de que o i-ésimo elemento é um objeto representativo de um dos grupos, $i = 1, 2, ..., n$ e z_{ij} como a variável indicadora de que o j-ésimo elemento foi designado ao grupo cujo medoide é o i-ésimo elemento, $i, j = 1, 2, ..., n$. Então é necessário resolver o seguinte problema de otimização, inicialmente proposto por Vinod (1969):

$$\text{minimizar} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} d(i, j)z_{ij}$$

sujeito a

$$\sum_{i=1}^{n} z_{ij} = 1, \quad j = 1, 2, ..., n$$

$$z_{ij} \leq y_i, \quad i, j = 1, 2, ..., n$$

$$\sum_{i=1}^{n} y_i = k$$

$$y_i, z_{ij} \in \{0, 1\}, \quad i, j = 1, 2, ..., n.$$

As condições garantem que exista um total de k $clusters$, sendo que cada um deles possui exatamente um elemento representativo. Ademais, cada elemento deve estar em exatamente um deles e garante-se que isso só pode ocorrer se ele foi atribuído ao respectivo grupo.

Outra definição do método é apresentada por Van der Laan et al. (2003). Nela, considera-se a matriz de dissimilaridades $\boldsymbol{D} = (d(i,j))_{n\times n} = (d(\boldsymbol{x}_i, \boldsymbol{x}_j))_{n\times n}$ simétrica e define-se $\boldsymbol{M} = (M_1, M_2, ..., M_k)$ qualquer coleção de k elementos de $\boldsymbol{X}$. Sendo $\boldsymbol{M}$ conhecido, pode-se calcular $d(\boldsymbol{x}_i, M_K)$ para cada $M_K \in \boldsymbol{M}$ e $i = 1, 2, ..., n$.

Dessa forma, sendo $\min_{K=1,2,...,k} d(\boldsymbol{x}_i, M_K) = d_1(\boldsymbol{x}_i, \boldsymbol{M})$ e $\min^{-1}_{K=1,2,...,k} d(\boldsymbol{x}_i, M_K) = l_1(\boldsymbol{x}_i, \boldsymbol{M})$, respectivamente, o mínimo e o minimizador de $d(\boldsymbol{x}_i, M_K)$, são selecionados os medoides $\boldsymbol{M^*} = \min^{-1}_M \sum_{i=1}^n d_1(\boldsymbol{x}_i, M)$. Assim, cada medoide M^*_K identifica um grupo, definido como todos os elementos que estão mais próximos a ele do que de qualquer outro medoide. Tal agrupamento é registrado nas etiquetas $l(\boldsymbol{X}, \boldsymbol{M}^*) = (l_1(\boldsymbol{x_1}, \boldsymbol{M}^*), ..., l_1(\boldsymbol{x_n}, \boldsymbol{M}^*))$.

2.1.2 Algumas medidas de dissimilaridade

Inicialmente, foi mencionado que esta técnica multivariada visa à minimização da soma das "distâncias" das observações aos medoides. Tais "distâncias" correspondem às dissimilaridades entre os elementos $\boldsymbol{x_i}$ e $\boldsymbol{x_j}$ $d(i,j)$, $i, j = 1, 2, ..., n$. Dessa forma, o método é bastante flexível, tendo em vista que pode ser facilmente adaptado para diversas funções que medem essa diferença ou outras que medem correlação.

Everitt et al. (2011) aponta que há dois principais tipos de medidas de dissimilaridade: distância e correlação. Definindo-se um peso w_k, $k = 1, 2, ..., p$ para a k-ésima variável quantitativa observada, destacam-se as seguintes:

- D1: distância euclidiana, dada por: $d(i,j) = \left(\sum_{k=1}^p w_k^2 (x_{ik} - x_{jk})^2\right)^{1/2}$;
- D2: distância Manhattan, dada por: $d(i,j) = \sum_{k=1}^p w_k |x_{ik} - x_{jk}|$;
- D3: distância de Minkowski, dada por: $d(i,j) = \left(\sum_{k=1}^p w_k^r |x_{ik} - x_{jk}|^r\right)^{1/r}$, $r \geq 1$;
- D4: distância de Camberra, dada por: $d(i,j) = 0$ se $x_{ik} = x_{jk} = 0$ ou $d(i,j) = \sum_{k=1}^p w_k \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}$;
- D5: correlação de Pearson, dada por: $d(i,j) = \frac{1-\phi_{ij}}{2}$, onde $\phi_{ij} = \frac{\sum_{k=1}^p w_k (x_{ik} - \overline{x}_{i\bullet})(x_{jk} - \overline{x}_{j\bullet})}{\left(\sum_{k=1}^p w_k (x_{ik} - \overline{x}_{i\bullet})^2 \sum_{k=1}^p w_k (x_{jk} - \overline{x}_{j\bullet})^2\right)^{1/2}}$, sendo $\overline{x}_{a\bullet} = \frac{\sum_{k=1}^p w_k x_{ak}}{\sum_{k=1}^p w_k}$, com $a = i, j$;
- D6: separação angular, dada por: $d(i,j) = \frac{1-\phi_{ij}}{2}$, onde $\phi_{ij} = \frac{\sum_{k=1}^p w_k x_{ik} x_{jk}}{\left(\sum_{k=1}^p w_k x_{ik}^2 \sum_{k=1}^p w_k x_{jk}^2\right)^{1/2}}$

Existem outras medidas de proximidade de observações multivariadas com características contínuas e/ou categóricas, como a proposta por Gower (1971). Mais detalhes podem ser encontrados em Everitt et al. (2011), Estabrook e Rogers (1966), entre outros.

2.1.3 Determinando o número de grupos

A fim de utilizar o PAM, é necessário determinar o número de grupos *a priori*. Dessa forma, foram desenvolvidas técnicas para estimar a quantidade de agrupamentos que devem ser utilizados. Dentre elas, Everitt et al. (2011) destaca o gráfico de silhuetas e as estatísticas GAP.

A primeira delas é descrita em Kaufman e Rousseeuw (2009) e cria gráficos para cada valor de k. Para cada $\boldsymbol{x_i}$, $i = 1, ..., n$, seja A o grupo a que ele foi atribuído pelo método PAM. Então, calcula-se a dissimilaridade média de $\boldsymbol{x_i}$ a todos os outros elementos de A, $a(\boldsymbol{x_i})$, (supõe-se que A tenha pelo menos dois elementos) e para cada um dos outros grupos C, $C \neq A$, computa-se $d(\boldsymbol{x_i}, C)$, análoga a $a(\boldsymbol{x_i})$, mas considerando apenas os elementos de C.

Na sequência, determina-se $b(\boldsymbol{x_i}) = \min_C d(\boldsymbol{x_i}, C)$, sendo B o grupo que realiza essa minimização. A silhueta de $\boldsymbol{x_i}$ é $s(\boldsymbol{x_i})$, definida como sendo:

$$s(\boldsymbol{x_i}) = \left(1 - \frac{a(\boldsymbol{x_i})}{b(\boldsymbol{x_i})}\right) I(a(\boldsymbol{x_i}) < b(\boldsymbol{x_i})) + \left(\frac{b(\boldsymbol{x_i})}{a(\boldsymbol{x_i})} - 1\right) I(a(\boldsymbol{x_i}) > b(\boldsymbol{x_i})) = \frac{b(\boldsymbol{x_i}) - a(\boldsymbol{x_i})}{\max\{a(\boldsymbol{x_i}), b(\boldsymbol{x_i})\}},$$

com $I(.)$ sendo a variável indicadora de $(.)$.

É possível mostrar que $-1 \leq s(\boldsymbol{x_i}) \leq 1$ para cada $\boldsymbol{x_i}$. Assim, $s(\boldsymbol{x_i})$ mede o quão bem $\boldsymbol{x_i}$ foi classificado e, dessa forma, quanto maior seu valor, mais parece que tenha ocorrido uma classificação correta. O gráfico de silhuetas é construído com os valores de $s(\boldsymbol{x_i})$ ordenados de forma decrescente em cada grupo e dispostos sequencialmente na horizontal. No eixo y, ficam os grupos e, dessa forma, quanto maior a 'altura' do gráfico de um agrupamento, mais elementos pertencem a ele.

Além disso, pode-se calcular a média das $s(\boldsymbol{x_i})$ de cada $cluster$ e a média geral de todas as silhuetas $\overline{s}(k) = \sum_{i=1}^{n} s(\boldsymbol{x_i})/n$. Este último valor pode ser utilizado para escolher o melhor k, isto é, escolher k que maximiza $\overline{s}(k)$. Em relação a uma possível interpretação de $\overline{s}(k)$, Kaufman e Rousseeuw (2009) apresentam uma tabela, a qual afirma que se $\overline{s}(k) \geq 0,71$, então foi encontrada uma estrutura forte de agrupamento, enquanto $0,51 \leq \overline{s}(k) \leq 0,70$ indica uma estrutura razoável.

Já as estatísticas GAP, propostas por Tibshirani et al. (2001), são uma maneira de formalizar a busca por um 'cotovelo' no gráfico de critério de agrupamento otimizado pelo número de grupos. Supondo que os dados foram classificados nos grupos $C_1, C_2, ..., C_k$, os quais contém $n_1, n_2, ..., n_k$ elementos, respectivamente, calcula-se

$$D_r = \sum_{\boldsymbol{x_i}, \boldsymbol{x_j} \in C_r} d(\boldsymbol{x_i}, \boldsymbol{x_j})$$

para cada $r = 1, 2, ..., k$ e define-se

$$W_k = \sum_{r=1}^{k} \frac{D_r}{2n_r}.$$

Dessa forma, busca-se padronizar o gráfico de $\log(W_k)$ ao compará-lo com seu valor esperado sob uma determinada distribuição de referência. Definindo

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

com $E_n^*(.)$ sendo a esperança de uma amostra de tamanho n sob a distribuição de referência. Então a estimativa do número de $clusters$ é $\hat{k}$, que maximiza $Gap_n(k)$.

2.1.4 Um algoritmo para a realização do PAM

Kaufman e Rousseeuw (2009) apresentam um algoritmo para a realização do Particionamento em Medoides descrito anteriormente. Ele é dividido em duas etapas: construção e troca; na primeira, são selecionados k objetos representativos; na outra, esses objetos são trocados até que a otimização seja realizada.

Assim, a etapa de construção inicialmente seleciona a observação com menor soma de dissimilaridades. Na sequência, são selecionadas outras, de modo que cada uma minimize a função objetivo, até que haja um total de k medoides. Para a seleção de cada um desses elementos, são seguidos os seguintes passos:

1. Considerar um objeto $\boldsymbol{x_i}$ que ainda não tenha sido selecionado;
2. Considerar outro objeto $\boldsymbol{x_j}$ que também não tenha sido selecionado e computar a diferença entre a sua dissimilaridade com o objeto mais semelhante já selecionado D_j e sua dissimilaridade com $\boldsymbol{x_i}$, denotada por $d(\boldsymbol{x_i}, \boldsymbol{x_j})$;
3. Se essa diferença for positiva, então $\boldsymbol{x_j}$ contribui para a seleção de $\boldsymbol{x_i}$ como elemento e calcula-se $C_{ji} = \max\{D_j - d(\boldsymbol{x_i}, \boldsymbol{x_j}), 0\}$;
4. Calcular o ganho total obtido por incluir $\boldsymbol{x_i}$ como medoide $\sum_j C_{ji}$;
5. Selecionar $\boldsymbol{x_i}$ que maximiza $\sum_j C_{ji}$ como medoide.

Na etapa de troca, tenta-se melhorar o conjunto de medoides selecionados a partir da comparação de todos os pares de objetos ($\boldsymbol{x_i}$,$\boldsymbol{x_j}$), onde $\boldsymbol{x_i}$ o primeiro foi selecionado, mas o segundo não. Para isso, Kaufman e Rousseeuw (2009) apontam os seguintes passos, que devem ser realizados para cada dupla:

1. Considerar um objeto não selecionado $\boldsymbol{x_h}$ e calcular sua contribuição C_{hij} para a troca de $\boldsymbol{x_i}$ por $\boldsymbol{x_j}$, dada por zero se estiver mais próxima de outro medoide do que de $\boldsymbol{x_i}$ ou $\boldsymbol{x_j}$. Caso contrário, se $d(\boldsymbol{x_i}, \boldsymbol{x_h}) = D_h$, define-se E_h como a dissimilaridade entre $\boldsymbol{x_h}$ e o segundo medoide mais próximo, e

   $C_{hij} = (d(\boldsymbol{x_h}, \boldsymbol{x_j}) - d(\boldsymbol{x_h}, \boldsymbol{x_i}))I(E_h > d(\boldsymbol{x_j}, \boldsymbol{x_h})) + (E_h - D_h)I(E_h \leq d(\boldsymbol{x_j}, \boldsymbol{x_h}))$.

   Caso contrário, $C_{hij} = d(\boldsymbol{x_h}, \boldsymbol{x_j}) - D_h$;

2. Calcular a soma das contribuições para a troca de $\boldsymbol{x_i}$ por $\boldsymbol{x_j}$, $T_{ij} = \sum_h C_{hij}$;
3. Selecionar o par ($\boldsymbol{x_i}$, $\boldsymbol{x_j}$) que minimiza T_{ij} ;
4. Se o mínimo de $T_{ij} < 0$, troca-se $\boldsymbol{x_i}$ por $\boldsymbol{x_j}$. Se não, o algoritmo para.

2.2 Considerações a respeito do Particionamento em Medoides

Swarndeep Saket e Pandya (2016) apontam que a facilidade de compreensão e implementação estão entre as vantagens deste método. De fato, ele se baseia em uma ideia intuitiva de buscar os melhores representantes dos grupos nos quais deseja-se agrupar os dados. Adicionalmente, Kaufman e Rousseeuw (2009) apresentam um algoritmo simples que realiza tal atribuição.

Além disso, eles também afirmam que ele é rápido e converge em um número finito de passos, o que é desejado de uma técnica de classificação. Entretanto, apesar dessa qualidade, ele não é escalável para bancos de dados grandes (isto é, com n elevado), o que inviabiliza seu uso nesses casos.

Ademais, por trabalhar com observações do banco de dados, é menos sensível a valores discrepantes do que outros métodos de classificação. Nesse sentido, é uma característica positiva, tendo em vista que torna o agrupamento realizado mais robusto a observações atípicas e $outliers$.

Por fim, possui maior custo computacional do que outras técnicas de agrupamento e, pelo algoritmo utilizado, o tempo utilizado depende das partições iniciais. Dessa forma, mesmo sendo menos sensível a valores extremos, há uma demora maior para a otimização dos medoides escolhidos.

2.3 Técnicas de agrupamento relacionadas ao Particionamento em Medoides

Uma lista de técnicas de agrupamento de dados multivariados pode ser encontrada em Saxena et al. (2017). Dentre elas, destacam-se BIRCH, ROCK, K-means e CLARA, as quais serão discutidas brevemente a seguir.

A primeira delas é a Redução Iterativa Balanceada e Agrupamento Usando Hierarquias (do inglês, BIRCH) e é utilizada quando se possui dados hierárquicos. Ela é baseada em características do grupo, isto é, armazena apenas o número de elementos, a soma linear e a soma quadrática (nesse sentido, assim como o PAM, só pode ser utilizado com dados contínuos). Além disso, é bastante útil para bancos de dados com muitas observações e é robusto a outliers.

Por outro lado, o ROCK pode ser aplicado quando há dados categóricos. Esse método é do tipo hierárquico aglomerativo e baseia-se no número de atributos semelhantes entre duas observações para a determinação dos grupos. Pela natureza das informações, nenhuma função relacionada a distância é utilizada.

Já a técnica K-means é adequada para agrupamentos via particionamento e é bastante semelhante ao PAM. A principal diferença reside no fato de que ela busca K centroides (não necessariamente vetores observados), sendo um para cada grupo, enquanto o particionamento via medoides procura K observações para minimizar a distância. Além disso, o algoritmo utilizado para K-means é mais rápido, o que torna-a preferível em alguns casos.

Por fim, segundo Kaufman e Rousseeuw (2009), o método CLARA é uma adaptação do PAM para situações em que há muitas observações. Dessa forma, são retiradas várias amostras e, em cada uma delas, é aplicado o PAM. Na sequência, são selecionados os K medoides que realizaram o melhor agrupamento dentre todas as amostras.

3 Aplicação em um conjunto de dados

O banco de dados que será utilizado para ilustrar o método está disponível em Wang (2020). Ele contém atributos de 178 vinhos produzidos na Itália, advindos de três classes diferentes. Entretanto, deseja-se desconsiderar esse número de grupos e realizar um novo agrupamento em K $clusters$, onde K deve ser determinado, para a construção de linhas de vinhos.

Para cada vinho, foram coletados 11 atributos numéricos contínuos, os quais deverão ser utilizados no agrupamento, a saber:

- Alcohol: graduação alcoólica do vinho;
- Malic acid: quantidade de ácido málico presente no vinho;
- Magnesium: quantidade de magnésio presente no vinho;
- Total phenols: quantidade total de fenóis presente no vinho;
- Flavanoids: quantidade de flavonoides presente no vinho;
- Nonflavanoid phenols: quantidade de fenóis não-flavonoides presente no vinho;
- Proanthocyanins: quantidade de proantocianinas presente no vinho;
- Color intensity: intensidade da cor do vinho;
- Hue: matiz do vinho
- OD280/OD315 of diluted wines: OD280/OD315 de vinhos diluídos;
- Proline: quantidade de prolina no vinho.

Com base nessas variáveis, foi estimado o número adequado de grupos de vinhos e, posteriormente, foi realizado o particionamento nesse número de medoides.

3.1 Determinação do número de agrupamentos

Para isso, foi utilizada a abordagem proposta por Kaufman e Rousseeuw (2009), que se baseia na aplicação sequencial do PAM para vários valores de k, seguido do cálculo de $\overline{s}(k)$ para cada K. Por fim, eles são ordenados e é selecionado como número de agrupamentos o K que maximiza $\overline{s}(k)$.

A Tabela 1 apresenta os resultados da aplicação dessa técnica, onde as linhas estão ordenadas em ordem decrescente de $\overline{s}(k)$. A partir dela, conclui-se que o número ideal de grupos é 2, que possui $\overline{s}(k) = 0,65$, o que indica uma razoável separação em grupos. Além disso, os agrupamentos gerados para outros valores são bastante inferiores, indicando que $K = 2$ é, de fato, um valor bastante razoável.

Tabela 1: Número de grupos (k) em ordem decrescente de média das silhuetas ($\overline{s}(k)$)

Número de grupos (k)	Média das silhuetas ($\overline{s}(k)$)
2	0,6496
3	0,5665
7	0,5637
4	0,5627
5	0,5472
6	0,5429

3.2 Aplicação do PAM para o melhor número de agrupamentos

Definido o número de grupos, aplicou-se a técnica, de forma a selecionar os dois medoides que representam os $clusters$. As observações que foram selecionadas como representantes foram as das linhas 46 e 144. A Figura 1 apresenta suas faces de Chernoff e, a partir dela, nota-se que, de fato, elas são bastante diversas.

Figura 1: Faces de Chernoff dos dois medoides

Denotando por 1 e 2 o grupo cujo medoide é, respectivamente, a observação 46 e 144, é possível calcular algumas estatísticas que resumem o agrupamento realizado. A Tabela 2 apresenta esses valores, de onde se conclui que o segundo grupo é maior do que o primeiro, com quase o dobro de elementos, além de que a dissimilaridade média do grupo 1 é superior, o qual também possui um máximo muito superior ao outro.

Tabela 2: Resumo do agrupamento realizado

Grupo	Número de elementos	Dissimilaridade máxima	Dissimilaridade média
1	62	600,02	165,64
2	116	272,11	113,13

As faces de Chernoff dos elementos atribuídos ao grupo 1 estão na Figura 2, enquanto as do grupo 2 estão nas Figuras 3 e 4. É possível perceber que os elementos são mais parecidos dentro do próprio grupo, apesar de que exista diversidade neles.

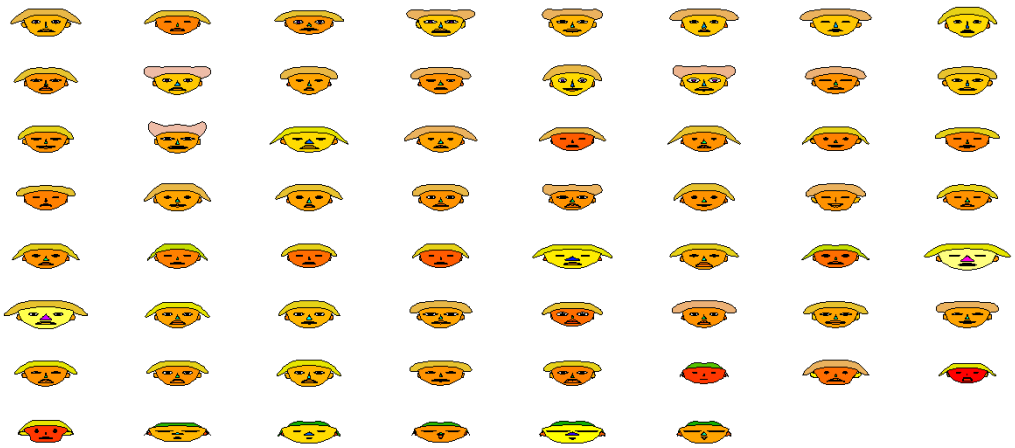

Figura 2: Faces de Chernoff dos elementos do primeiro grupo

Figura 3: Faces de Chernoff dos elementos do segundo grupo

Figura 4: Faces de Chernoff dos elementos do segundo grupo

3.3 Discussão da técnica aplicada

Em relação à técnica de agrupamento aplicada, ela foi adequada no sentido de solucionar o problema proposto inicialmente, isto é, determinar o número de grupos e atribuir os vinhos a cada um deles. Dessa forma, através de sua aplicação, foi possível dar uma resposta condizente para o cliente em questão.

Além disso, uma vantagem da utilização desta técnica em relação a outras semelhantes - sobretudo o K-médias - é o fato de retornar o elemento mais representativo de cada linha de vinho. Ou seja, como foram selecionados os dois medoides que minimizaram as dissimilaridades, foi possível apontar para o cliente um vinho típico que representa cada uma das linhas produzidas.

Por outro lado, caso seja de interesse, no futuro, repetir o processo com um número maior de observações, isso poderia se tornar inviável pelo alto custo computacional do método. Dessa forma, seria melhor optar por métodos de complexidade mais baixa ou adaptados para tais situações, como as K-médias ou CLARA.

A respeito de possíveis melhorias, como não há informações de como foi realizada a coleta dos dados, poderia ter sido elaborado um planejamento estatístico de experimentos, de modo a garantir que os agrupamentos sejam, de fato, adequados. Isto é, não se sabe se todos foram analisados pela mesma pessoa, em qual ordem, sob quais condições de temperatura, dentre outras características que podem interferir nas medidas e, portanto, no agrupamento realizado.

4 Conclusão

Com base na definição e discussão realizada acerca do Particionamento em Medoides (PAM), pode-se concluir que é uma técnica de agrupamento bastante simples e que permite interpretação. Dessa forma, apesar de suas limitações, pode ser utilizada em vários contextos a fim de se obter agrupamentos e elementos representativos de cada um deles.

Além disso, a partir da aplicação apresentada, nota-se que foi encontrada uma estrutura de grupos razoável e que permitiu o agrupamento dos vinhos produzidos de uma forma que facilita a comercialização dos mesmos. Assim, a utilização desse método levou a respostas satisfatórias às questões iniciais.

Referências

Estabrook, G. F. e Rogers, D. J. (1966). A general method of taxonomic description for a computed similarity measure. *BioScience*, 16(11):789–793.

Everitt, B., Landau, S., Leese, M., e Stahl, D. (2011). Cluster analysis.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

Kaufman, L. e Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., e Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267:664–681.

Swarndeep Saket, J. e Pandya, S. (2016). An overview of partitioning algorithms in clustering techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5(6):1943–1946.

Tibshirani, R., Walther, G., e Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Van der Laan, M., Pollard, K., e Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584.

Vinod, H. D. (1969). Integer programming and the theory of grouping. *Journal of the American Statistical association*, 64(326):506–519.

Wang, H. (2020). Wine dataset for clustering.