DOI: 10.48011/asba.v2i1.1194

Índice de validação de agrupamento de dados baseado em curvas principais

Franciele A. Ferreira * **Danton D. Ferreira** ** **Bruno H. G. Barbosa** ***

* *Departamento de Automática, Universidade Federal de Lavras, MG, (e-mail: franciele.ferreira1@posgrad.ufla.br).*
** *Departamento de Automática, Universidade Federal de Lavras, MG, (e-mail: danton@ufla.br).*
*** *Departamento de Automática, Universidade Federal de Lavras, MG, (e-mail: brunoh@ufla.br).*

Abstract: The Data clustering algorithms are unsupervised methods that seek to separate the data set into K groups without relying on prior knowledge of the data set. Cluster validation indices are commonly used to validate the result of these algorithms, since no label are available. The clustering of data based on Principal curves is a nonlinear generalization of Principal Component Analysis and can map the nonlinearities of data from various dimension to one dimension, generating thus compact data representation vectors. It is capable to deal with data clusters with non-Gaussian distribution, but in general, validation indices are constructed by combining a measure of compression and a measure of separation that does not always reflect the quality of the cluster for this type of data distribution and may thus present misleading results. In this article paper, a new validation index is proposed for data clustering based on principal curves. Experiments on synthetic data clusters were carried out. The results showed that the proposed index is suitable for the principal curves-based clustering method for different data distributions.

Resumo: Os algoritmos de agrupamento de dados são métodos não supervisionados que buscam separar o conjunto conjunto de dados em k grupos sem depender do conhecimento prévio do conjunto de dados. Os índices de validação de *cluster* são comumente usados para validar o resultado desses algoritmos, já que nenhum rótulo está disponível. O agrupamento de dados com base nas curvas principais, é uma generalização não linear da Análise de Componentes Principais e pode mapear as não linearidades dos dados de várias dimensões para uma dimensão, gerando vetores de representação de dados compactos. Ele é capaz de lidar com *clusters* de dados com distribuição não gaussiana, mas, em geral, os índices de validação são construídos combinando uma medida de compactação e uma medida de separação que nem sempre reflete a qualidade do *cluster* para esse tipo de distribuição de dados e pode, portanto, apresentar resultados enganosos. Neste artigo, um novo índice de validação é proposto para agrupamento de dados com base nas curvas principais. Foram realizados experimentos em bases de dados sintéticas. Os resultados mostraram que o índice proposto é adequado para o método de agrupamento baseado em curvas principais para diferentes distribuições.

Keywords: Validation Indices; Data Clustering; Principal Curves; Unsupervised Method; Non-Gaussian Distribution.

Palavras-chaves: Índices de Validação; Agrupamento de Dados; Curvas Principais; Método Não Supervisionado; Distribuição Não Gaussiana.

1. INTRODUÇÃO

Os algoritmos de agrupamento de um modo geral buscam separar os dados em k grupos de acordo com o grau de similaridade entre eles. A presença de uma grande variabilidade nos tamanhos, formas e densidades dos agrupamentos e o número de *clusters* (grupos) que nem sempre são conhecidos a priori, são um dos grandes desafios do processo de agrupamento. Além disso, se tratando de um método não supervisionado, no agrupamento de dados não se tem um conhecimento prévio sobre o conjunto original e a validação desses agrupamentos acaba se tornando uma tarefa essencial.

Os índices de validação de agrupamento são comumente usados para realizar a validação das partições produzidas pelos algoritmos de agrupamento. Segundo De Castro and Ferrari (2016), os índices de validação são responsáveis por aferir de forma quantitativa o agrupamento de dados resultante de um algoritmo para uma base de dados. Assim, o particionamento que otimizar melhor o índice em questão é selecionado como resultado final.

De acordo com Zalik and Zalik (2011) a maioria dos índices de validação propostos nas últimas décadas se concentrou na compactação e na separação dos dados. A medida de compactação mede quão próximos estão os objetos pertencentes a um mesmo *cluster* e a medida de separação mede quão separados estão os *clusters*.

Uma medida clássica de compactação é a variação. Um baixo valor de variação indica que o *cluster* é mais compacto. Quanto à separação, geralmente é utilizada a distância entre os centros dos *clusters*. No entanto, essa distância nem sempre reflete a qualidade da partição e às vezes, fornece resultados incorretos. Segundo Zalik and Zalik (2011), as medidas de validação mais populares têm a tendência de ignorar *clusters* com baixa densidade e não são eficientes na validação de partições com tamanhos e densidades diferentes.

Existem alguns algoritmos na literatura como DBSCAN (Ester et al., 1996), Kernel K-means (Dhillon et al., 2004) e agrupamento baseado em curvas principais (Ferreira et al., 2019) que tentam resolver o problema de *clusters* não linearmente separáveis, com diferentes tamanhos e formas. Os índices clássicos de validade de agrupamento acabam não suportando esses tipos de distribuição, podendo apresentar resultados que não representam a realidade do conjunto de dados.

Neste trabalho é proposto um índice de validação de agrupamento inspirado no índice *silhouettes* (Rousseeuw, 1987) para validar o resultado do particionamento gerado pelo algoritmo de agrupamento iterativo baseado em curvas principais (Ferreira et al., 2019).

As curvas principais tem sido utilizadas em vários trabalhos como no trabalho de Moraes et al. (2019) onde foi proposto um método de agrupamento de dados não iterativo que utiliza o algoritmo k-segmentos para construção da curva principal. Já Braga et al. (2018) propuseram um método de seleção automática de parâmetros, utilizando o algorítimo TLBO, para o algoritmo k-segmentos, que constrói curvas principais. No trabalho de Ferreira et al. (2013) foi feita um investigação sobre como as curvas principais podem ser aplicadas ao problema de monitoramento da qualidade de energia em termos de análise, detecção e classificação de eventos.

O índice proposto também se baseia nas medidas de compactação e separação, porém leva em consideração o formato dos *clusters* através das curvas principais pelas quais os mesmos são representados a partir do algoritmo baseado em curvas principais. O índice foi aplicado a quatro bases de dados sintéticas com diferentes distribuições e os resultados foram comparados ao índice *Silhouettes* (Rousseeuw, 1987).

O restante do trabalho está organizado da seguinte maneira. Na seção 2 é feita uma contextualização sobre os índices de validação de agrupamento. Na subseção 2.1 é demonstrado como é feito a validação de *clusters*, na seção 2.2 é descrito o funcionamento do algoritmo de agrupamento baseado em curvas principais e na subseção 2.3 é apresentado o índice *silhouettes*. Já na seção 3 é explicado o índice proposto e por fim nas seções 4 e 5 são apresentados os resultados experimentais e a conclusão respectivamente.

2. ÍNDICES DE VALIDAÇÃO DE AGRUPAMENTO

De acordo com Jain and Dubes (1988), os índices de validação podem ser classificados em externos, internos e relativos.

Segundo De Castro and Ferrari (2016), os índices externos são medidas que avaliam o quão correto está um agrupamento de dados em relação a um agrupamento de referência ideal que se deseja alcançar. O cálculo dessas medidas requer o conhecimento prévio do grupo ao qual cada elemento pertence.

De Castro and Ferrari (2016) citam que os índice internos medem o grau em que uma solução de agrupamento é justificada com base apenas nas informações intrínsecas aos elementos do conjunto de dados original ou em uma matriz de similaridades ou dissimilaridades calculadas a partir do mesmo.

Já em relação aos índices relativos, de acordo com Jain and Dubes (1988), os mesmos medem a qualidade relativa entre dois agrupamentos diferentes, servindo para compará-los.

2.1 Validação de Cluster

A validação de *cluster* é um procedimento que busca avaliar os resultados de um algoritmo de agrupamento sendo realizado por índices de validação. Segundo Wani and Riyaz (2016) a validade de *cluster* ajuda também a resolver tanto o problema de definição de número de *clusters* (nc) e definição dos centróides iniciais dos *clusters*.

De acordo com Wani and Riyaz (2016) e Zhao (2014), para encontrar um número apropriado de *clusters* considerando um conjunto de dados $\boldsymbol{X}$, um algoritmo de agrupamento específico e uma faixa de número de grupos, é necessário realizar os seguintes passos:

(1) Repetir um algoritmo de agrupamento sucessivas vezes de acordo com um número de grupos definidos em uma faixa fixada de valores: $nc \in [nc_{min}, nc_{max}]$;
(2) Obter o resultado do agrupamento para cada valor de nc definido na faixa;
(3) Calcular o valor do índice de validade para todas as soluções;
(4) Selecionar o nc, valor máximo ou mínimo (Figura 1), para o qual o particionamento de dados forneceu o melhor resultado.

2.2 Agrupamento de dados baseado em Curvas Principais

Segundo Ferreira et al. (2019) o método de agrupamento baseado em curvas principais, particiona o conjuntos de dados em k *clusters*, alocando curvas principais (Hastie and Stuetzle, 1989) em cada um desses particionamentos. Em um segundo passo, de forma iterativa, o método calcula a distância de cada evento às curvas alocadas e associa cada um à curva que estiver mais próximo. Havendo mudança de algum evento em relação à curva que ele estava associado anteriormente, as curvas são redefinidas e todo procedimento é repetido até que não haja mudança.

De acordo com Ferreira et al. (2019), as curvas principais são compostas por um conjunto de segmentos gerados pelo

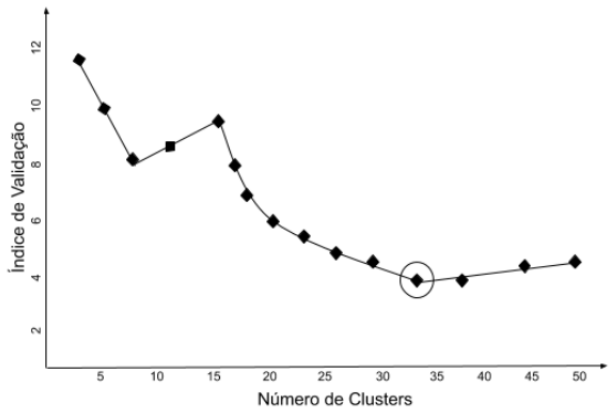


Figura 1. Seleção nc de acordo com o índice de validação

algoritmo k-segmentos (Verbeek et al., 2002). Segundo os autores, as curvas principais conseguem descrever bem o formato dos *clusters* fazendo uso da sua alta capacidade de representação unidimensional de um conjunto de dados multidimensional. Assim, um conjunto de dados com *clusters* de diferentes distribuições e formatos conseguem ser melhor particionados.

As etapas do método de agrupamento de dados baseado em curvas principais para particionar o conjunto de dados são descritas a seguir:

(1) Definição dos *clusters* iniciais por meio do algoritmo k-means.
(2) Alocação de curvas, geradas a partir do algoritmo k-segmentos, em cada um dos *clusters* gerados.
(3) Cálculo da distância de cada evento às curvas de todos os *clusters*.
(4) Associação do evento ao *cluster* correspondente à curva que ele está mais próximo.
(5) Calculo de novas curvas para cada um dos *clusters* caso algum evento altere de *cluster*.
(6) Repetição dos passos de 2 à 5 até a convergência, ou seja, quando nenhum evento trocar de *cluster*.

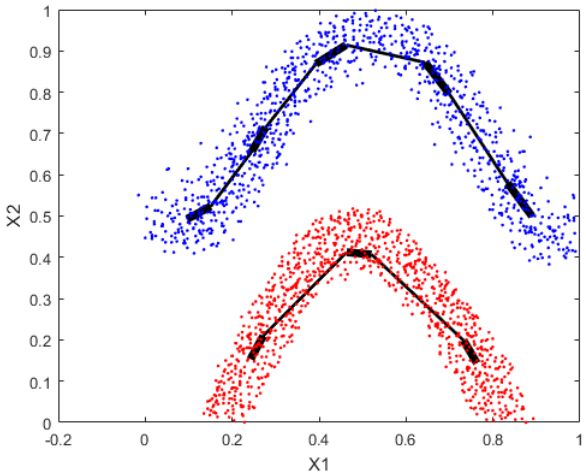


Figura 2. Resultado de agrupamento de dados baseado em curvas principais

A Figura 2 mostra o resultado do particionamento realizado pelo algoritmo de agrupamento baseado em curvas principais. O conjunto de dados possui dois *clusters* sendo um na cor azul e outra na cor vermelha. É possível observar as curvas que o algoritmo*k-segmentos* gerou representando cada um dos *clusters* que foram identificados. As linhas mais grossas representam os segmentos e as mais finas representam a ligação entre esses segmentos.

2.3 Índice silhoutte

O índice de *Silhouettes* proposto por Rousseeuw (1987) avalia a qualidade de um agrupamento de dados com base nas distâncias *inter-cluster*, que mede a separação dos grupos e *intra-cluster*, que mede a compactação.

Dado um conjunto de dados $X = \{x_1, x_2, ..., x_n\}$ e o resultado do seu agrupamento $C = \{C_1, ..., C_k\}$, considerando NC_i e NC_k o número de elementos dos *clusters* C_i e C_k e sendo *d(i,j)* a distância entre os pontos i e j no *cluster* C_i, o valor silhueta *S(i)* pode ser calculado conforme mostrado em (1):

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, se|NC_i| > 1 \tag{1}$$

onde,

$$a(i) = \frac{1}{NC_i - 1} \sum_{j \in C_i, i \neq j} d(i,j) \tag{2}$$

$$b(i) = \min_{k \neq i} \frac{1}{NC_k} \sum_{j \in C_k} d(i,j) \tag{3}$$

De acordo com Batool and Hennig (2019) *a(i)* é a distância média de i aos pontos no *cluster* ao qual foi atribuído e *b(i)* é a distância média de i aos pontos do *cluster* mais próximo ao qual não foi atribuído. Assim um grande valor de *S(i)* indica que i está muito mais próximo das observações em seu próprio *cluster* do que no vizinho.

Após o cálculo do valor de silhueta para cada objeto i, a silhueta média global pode ser calculada para avaliar a qualidade do agrupamento em relação à todo conjunto de objetos. A silhueta média global pode ser obtida conforme é apresentado em (4), sendo N a quantidade de elementos do agrupamento:

$$S = \frac{1}{N} \sum_{i=1}^{N} s(i) \tag{4}$$

A Figura 3 ilustra os elementos envolvidos no cálculo do valor de silhueta do objeto i pertencente ao *cluster* C_1. Cada linha que liga o objeto i ao outros objetos do *cluster* C_1 representa as distâncias *intra-clusters* usadas para calcular a parte *a(i)* do cálculo do índice *silhouettes*.

Já as linhas que ligam o objeto i aos objetos dos *clusters* C_2 e C_3 representam as distâncias *inter-cluster*. Para o cálculo da parte *b(i)* podemos observar que serão utilizadas as distâncias *inter-clusters* referente ao *cluster* C_3 visto que o mesmo se encontra mais próximo ao C_1.

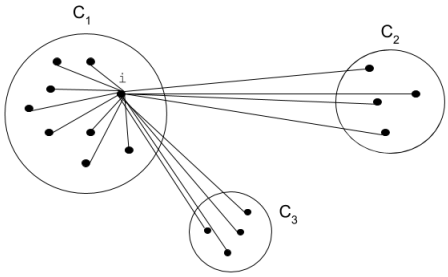


Figura 3. Elementos envolvidos no cálculo de *S(i)*

3. ÍNDICE DE VALIDAÇÃO PROPOSTO

O índice de validação proposto, *Silhouettes for Principal Curvers* - SPC, estabelece uma nova medida de compactação que faz uso das curvas geradas pelo algoritmo de agrupamento com curvas principais (Ferreira et al., 2019).

Inspirando no índice *silhouettes*, são calculadas as medidas *intra-cluster* (compactação) e *inter-cluster* (separação). Para o cálculo da medida de compactação é calculada a distância, *ac(i)*, do objeto i à curva principal do *cluster* a qual ele faz parte, c_i.

Em relação à medida de separação é calculada a distância, bc(i), do objeto i à curva principal referente ao *cluster* mais próximo do qual ele foi atribuído, c_k.

Comparando com o índice *silhoutte*, *S(i)*, ao invés de considerar a distância de i a cada objeto do próprio *cluster* ou do *cluster* mais próximo, foi considerada a distância às curvas principais de cada um dos grupos gerados visto que a curva principal o representa de forma unidimensional.

A equação para obtenção do valor do índice proposto é apresentado em 5.

$$SPC(i) = \frac{bc(i) - ac(i)}{max\{ac(i), bc(i)\}}, se|C_i| > 1 \tag{5}$$

Onde,

$$ac(i) = d_{min}(i, c_i) \tag{6}$$

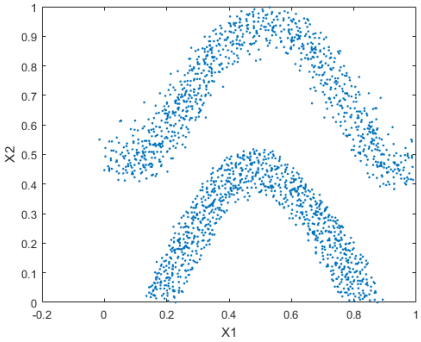
$$bc(i) = d_{min}(i, c_k) \tag{7}$$

4. BASE DE DADOS

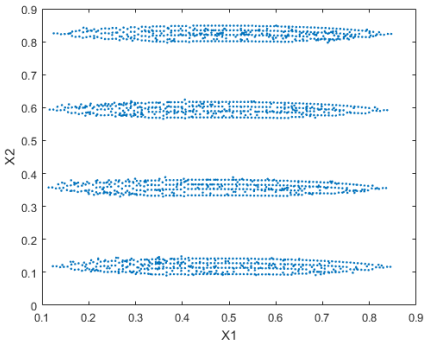
De modo a avaliar o índice proposto, foram realizados testes em quatro bases de dados sintéticas próprias, com formatos, tamanhos e distribuições diferentes. A quantidade de eventos por classe para cada uma das bases de dados é apresentado na Tabela 1. Vale ressaltar que as classes sinalizadas com *(-)* é um indicativo de que a base de dados não possui essa classe.

O método de agrupamento utilizado para gerar as partições que são parâmetros de entradas para os índices de validação foi o método baseado em curvas principais (Ferreira et al., 2019) e para fins de comparação, utilizou-se o índice *silhouettes* (Rousseeuw, 1987).

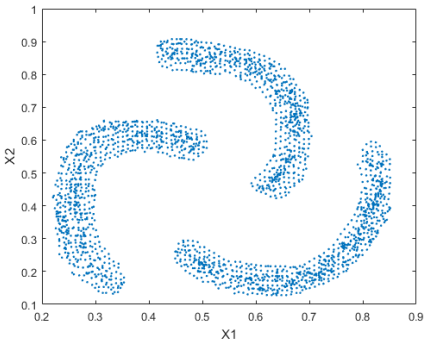
O formato de distribuição dos dados em cada uma das bases de dados é apresentado na Figura 4.



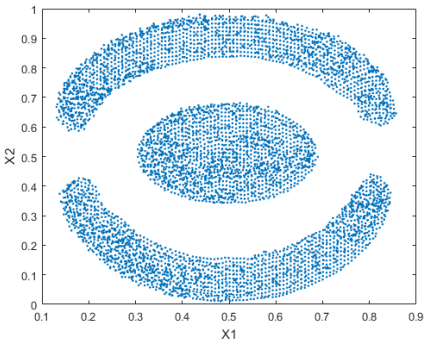
(a) *Base de Dados Parábolas*



(b) *Base de Dados Elipses Horizontais*



(c) *Base de Dados Espiral*



(d) *Base de Dados Olho*

Figura 4. Formato das Bases de Dados

O método de agrupamento baseado em curvas principais conta com dois parâmetros de entrada, k que representa o

Tabela 1. *Número de eventos por classe*

Base de Dados	A	B	C	D	Total
Parábolas	1000	1000	-	-	2000
Elipses Horizontais	440	440	440	440	1760
Espiral	800	780	680	-	2260
Olho	2110	2048	1840	-	5998

número de segmentos da curva principal e f que representa o comprimento de cada segmento.

A partir de vários testes com valores diferentes para o número de segmentos e tamanho de segmentos da curva principal, foi possível determinar um valor de k e f que melhor representou a curva principal para cada uma das bases de dados de acordo levando em consideração a maior acurácia obtida.

A configuração dos parâmetros k e f bem como a acurácia obtida pelo método de agrupamento baseado em curvas principais podem ser visto na Tabela 2.

conforme a Tabela 2.

Tabela 2. *Valor dos parâmetros k e f e Acurácia*

Base de Dados	k	f	Acurácia
Parábolas	5	0.4	0.9075
Elipses Horizontais	8	0.5	1
Espiral	8	0.5	1
Olho	10	0.5	1

5. RESULTADOS E DISCUSSÃO

Para avaliar o desempenho do índice de validação proposto, realizou-se experimentos em vários conjuntos de dados sintéticos com número conhecido de *clusters*. A metodologia clássica de avaliação consiste em aplicar um algoritmo de *clustering* em cada conjunto de dados, variando o número de *clusters* e calculando o índice de validação para cada partição onde se determina que o número real de *clusters* é o que melhor se ajusta aos dados.

As Tabelas 3, 4, 5 e 6 apresentam os valores do índice *silhouettes (S)* e do índice proposto *(SPC)*, alcançados em cada uma das bases de dados de acordo com a variação do número de *clusters*.

Tabela 3. Valores índices - Base Parábolas

Nº Clusters	S	SPC
2	0,5167	0,9651
3	0,4804	0,9524
4	0,5719	0,9443
5	0,5687	0,9363

Tabela 4. Valores índices - Base Elipses Horizontais

Nº Clusters	S	SPC
2	0,4934	0,9922
3	0,4873	0,9942
4	0,4787	0,9950
5	0,4048	0,9885

Já as Figuras 5, 6, 7 e 8 mostram de forma gráfica a relação dos valores dos índices *silhouettes* e *silhouettes pc* de acordo com a variação do número de *clusters* para cada uma das bases de dados.

Tendo conhecimento do número de classes iguais aos apresentados na Tabela 1 ao comparar os valores obtidos

Tabela 5. Valores índices - Base Espiral

Nº Clusters	S	SPC
2	0,3614	0,9769
3	0,5849	0,9921
4	0,5447	0,9860
5	0,5859	0,9816

Tabela 6. Relação valores índices - Base Olho

Nº Clusters	S	SPC
2	0,5514	0,9814
3	0,4317	0,9821
4	0,5416	0,9740
5	0,6593	0,9687

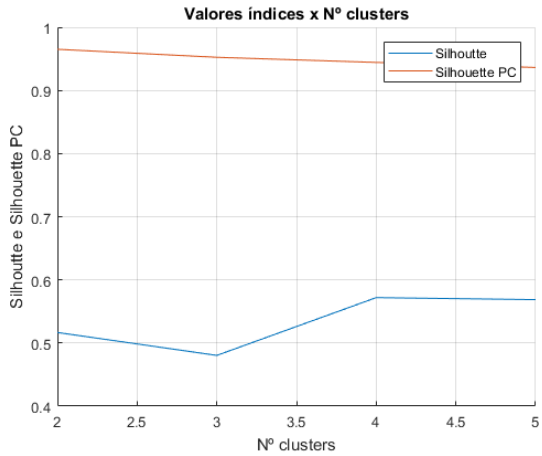


Figura 5. Índice Silhouette x Silhouette PC - Base Parábolas

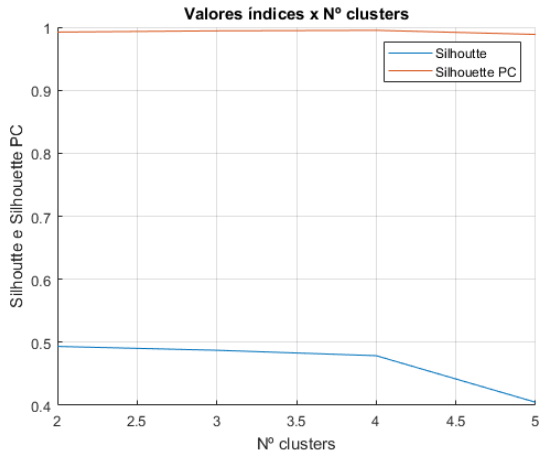


Figura 6. Índice Silhouette x Silhouette PC - Base Elipses Horizontal

pelo índice proposto e pelo índice *silhouettes* é possível observar que o índice proposto, para todos as bases de dados, apresentou um valor maior quando o número de *clusters* eram iguais aos números de classes esperada indicando que essa quantidade de *cluster* era a que melhor se ajustava os dados.

Já observando o comportamento do índice *silhouettes*, é possível verificar que para todas as base de dados o valor maior do índice indicava uma quantidade de *clusters* diferente do número de classe esperada.

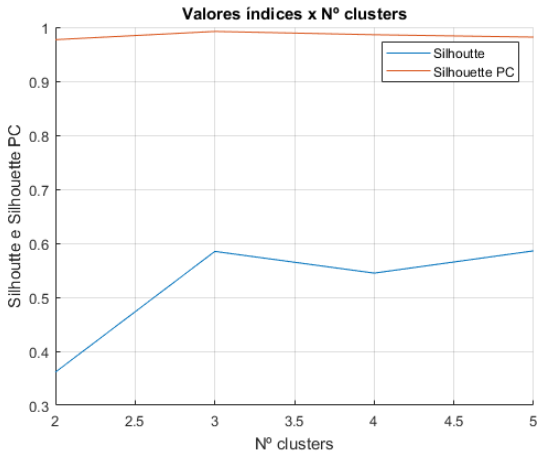


Figura 7. Índice Silhouette x Silhouette PC - Base Espiral

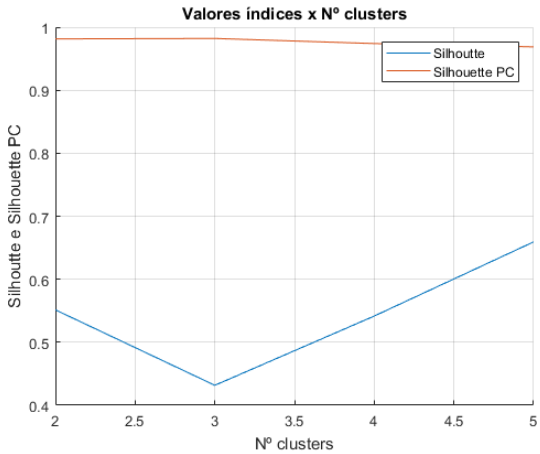


Figura 8. Índice Silhouette x Silhouette PC - Base Olho

Com essa análise é possível observar que o índice proposto conseguiu indicar uma quantidade de *clusters* iguais a quantidade de classes original para todas as bases de dados.

Uma outra vantagem do índice proposto está relacionada ao custo computacional menor que o do *silhouettes* padrão. O cálculo de distância de um elemento i a todos os outros elementos do *cluster* a qual ele pertence e ao *cluster* mais próximo é substituído apenas pelo cálculo da distância à curva a qual ele faz parte e da curva do *cluster* mais próximo evitando assim efetuar vários cálculos de distância para encontrar uma média.

Pode-se verificar que o índice proposto conseguiu refletir melhor o particionamento gerado pelo método de agrupamento baseado em curvas principais comparado ao índice *silhouettes*.

6. CONCLUSÃO

Apesar de existirem vários índices de validação de agrupamento na literatura, há uma necessidade de melhorar sua eficácia e robustez para avaliar agrupamentos de dados arbitrários, como dados não gaussianos e não linearmente separáveis. Neste trabalho, foi proposto um novo índice de validade para validar o particionamento realizado pelo algoritmo de agrupamento baseado em curvas principais.

A análise experimental mostra que o índice de validade de *cluster* proposto é eficaz para a estrutura de dados arbitrária. A abordagem proposta trata-se de uma solução simples e eficiente, baseado no índice *silhouetttes*, sem complexidade adicional.

A eficácia do índice proposto foi validada por quatro conjuntos de dados sintéticos com distribuição arbitrária onde o mesmo apresentou melhores resultados. A originalidade do trabalho desenvolvido inclui a proposta de um novo índice de validade de *cluster* para agrupamento de dados baseado em curvas principais.

AGRADECIMENTOS

Agradecemos a Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) e ao Consellho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio com a infraestrutura disponibilizada para o desenvolvimento da pesquisa.

REFERÊNCIAS

Batool, F. and Hennig, C. (2019). Clustering with the average silhouette width.

Braga, A.A., Ferreira, D.D., and Barbosa, B.H.G. (2018). Seleção automática de parâmetros iniciais do algoritmo k-segmentos com teaching-learning-based optimization. *XXII Congresso Brasileiro de Automática*, 1–7.

De Castro, L. and Ferrari, D. (2016). *Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações*.

Dhillon, I.S., Guan, Y., and Kulis, B. (2004). Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, 551–556. ACM, New York, NY, USA. doi: 10.1145/1014052.1014118. URL `http://doi.acm.org/10.1145/1014052.1014118`.

Ester, M., Kriegel, H.P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*.

Ferreira, D.D., de Seixas, J.M., Cerqueira, A.S., and Duque, C.A. (2013). Exploiting principal curves for power quality monitoring. *Electric Power Systems Research*, 100, 1 – 6.

Ferreira, F., Braga, A., Ferreira, D., and Barbosa, B. (2019). Método de agrupamento iterativo baseado em curvas principais. doi:10.17648/sbai-2019-111450.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406), 502–516.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. URL `http://portal.acm.org/citation.cfm?id=46712`. Attention: PDF is rather large (39MB).

Moraes, E.C.C., Ferreira, D.D., Vitor, G.B., and Barbosa, B.H.G. (2019). Data clustering based on principal curves. *Advances in Data Analysis and Classification*. doi:10.1007/s11634-019-00363-w. URL `https://doi.org/10.1007/s11634-019-00363-w`.

Rousseeuw, P. (1987). Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal*

of Computational and Applied Mathematics, 20, 53–65. doi:10.1016/0377-0427(87)90125-7.

Verbeek, J., Vlassis, N., and Kröse, B. (2002). A k-segments algorithm for finding principal curves. *Pattern Recognition Letters*, 23(8), 1009 – 1017.

Wani, M. and Riyaz, R. (2016). A new cluster validity index using maximum cluster spread based compactness measure. *International Journal of Intelligent Computing and Cybernetics*, 9, 179–204. doi:10.1108/IJICC-02-2016-0006.

Zalik, K. and Zalik, B. (2011). Validity index for clusters of different sizes and densities. *Pattern Recognition Letters*, 32, 221–234. doi:10.1016/j.patrec.2010.08.007.

Zhao, Q. (2014). Wb-index: A sum-of-squares based index for cluster validity. *Data & Knowledge Engineering*, 92. doi:10.1016/j.datak.2014.07.008.

.