Um Estudo de Variantes do Índice de Validação Silhueta

Victória Vargas[1], Eduardo Amorim[2], José André de M. Brito[1], Gustavo S. Semaan[3]

[1] Escola Nacional de Ciências Estatísticas (ENCE-IBGE), Rio de Janeiro, RJ, Brasil

[2]Universidade Anhanguera (Compus Niterói), Niterói, RJ, Brasil

[3]Universidade Federal Fluminense (INFES-UFF), Santo Antônio de Pádua, RJ, Brasil

victoriavargasestudo@gmail.com, jose.m.brito@ibge.gov.br

Abstract. *This paper aims to evaluate five variants of the silhouette index for their ability to detect good quality solutions to clustering problems. Five computational experiments were carried out, covering 51 diversified databases (natural and artificial). As dissimilarity measures, Euclidean and Manhattan distances were used, and for clustering algorithms PAM, DBSCAN, and Bisecting k-means. The results obtained indicate that the median-based variant is a good alternative to detect quality solutions.*

Resumo. *Este artigo se propõe a avaliar cinco variantes do índice de silhueta quanto à sua capacidade de detectar soluções de boa qualidade para problemas de agrupamento. Foram realizados cinco experimentos computacionais, contemplando 51 instâncias da literatura diversificadas (dados reais e artificiais). Como medidas de dissimilaridade foram utilizadas as distâncias euclidiana e de manhattan e para os algoritmos de agrupamento, PAM, DBSCAN e Bisecting k-means. Os resultados obtidos indicam que a variante baseada na mediana constitui-se como boa alternativa para detectar soluções de qualidade.*

1. Introdução

A análise de agrupamentos corresponde uma técnica de mineração de dados que abarca uma coleção de algoritmos para a resolução de problemas de agrupamento (PA). De acordo com [Han et al. 2012], dado um conjunto X constituído por n objetos, de forma que $X = \{x_1, x_2, ..., x_n\}$ e cada objeto x_i possui p atributos, resolver um problema de agrupamento consiste em construir, a partir de X, k grupos $G_r, r = 1, ..., k$ que definem uma solução $\Pi = \{G_1, ..., G_k\}$. Como pressuposto desse problema, os objetos alocados a um mesmo grupo devem ter baixo grau de dissimilaridade entre si com base em seus p atributos. Adicionalmente, a estrutura de grupos produzida (alocação dos objetos aos grupos) depende da medida de dissimilaridade e do algoritmo escolhido [Bussab et al. 1990].

Dada uma solução, é de suma importância verificar se ela corresponde a uma boa estrutura de agrupamento [Kaufman and Rousseeuw 1989]. Nesse sentido, este trabalho utilizou cinco versões do Índice de Validação Silhueta, e analisou soluções produzidas por algoritmos clássicos da literatura, sejam eles: PAM (*Partitioning Around Medoids*), DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) e Bisecting k-means (BK) [Han et al. 2012].

2. Metodologia

A metodologia utilizada neste trabalho contempla: (i) algoritmos para construção de soluções; (ii) distâncias capazes de quantificar a dissimilaridade entre pares de objetos; (iii) índice para mensurar a qualidade das soluções; (iv) instâncias (bases) da literatura com características diversificadas para a realização de experimentos computacionais.

Para a construção de soluções, este estudo utilizou os algoritmos PAM, DBSCAN e BK. O PAM e o BK possuem, como parâmetro de entrada, a quantidade (k) de grupos a ser formada. Já o DBSCAN tem seus parâmetros relacionados ao conceito de densidade (quantidade de objetos em uma dada região). Neste caso, uma abordagem para calibrar os parametros de entrada para o DBSCAN, intitulada DistK, foi considerada [Semaan 2013]. Em relação à dissimilaridade entre objetos foram utilizadas as Distâncias Euclidiana (DE) e de Manhattan (DM). Como principal alvo de estudo na pesquisa, a seção 3 apresenta cinco variantes do Índice Silhueta, utilizados com o intuito de avaliar a qualidade das soluções produzidas pelos algoritmos de agrupamento.

3. Índices Silhueta

Proposto por [Kaufman and Rousseeuw 1989], o índice Silhueta *Tradicional* (ST) é calculado da seguinte maneira: $a(x_i)$ corresponde à distância média de cada objeto x_i em relação aos demais objetos do mesmo grupo (distância *intracluster*), e $b(x_i)$ é a menor distância média de x_i aos demais grupos (distância *interclusters*). Utilizando-se os valores dessas distâncias, calcula-se o valor da silhueta de cada objeto equação (1), sendo a silhueta da solução correspondente à média aritmética dessas silhuetas.

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{b(x_i), a(x_i)\}} \qquad i = 1, \ldots, n \tag{1}$$

Também bem conhecida na Literatura, a Silhueta Simplificada (SS) apresenta resultados de qualidade comparáveis aos da Silhuta Tradicional, com a vantagem de demandar menor custo computacional. Para essa variante, $a(x_i)$ é a distância entre x_i e o centroide do mesmo grupo c_r ($x_i \in G_r$), enquanto $b(x_i)$ é a menor distância entre x_i e o centroide c_s de um outro grupo G_s ($r \neq s$) [Hruschka et al. 2004].

Propostas por [Amorim 2013], as Silhueta Alternativas 1 e 2 (SA1 e SA2) proporcionaram soluções equivalentes ou superiores às observadas na literatura, no que diz respeito aos experimentos conduzidos com o DBSCAN em um subconjunto de instâncias de DS2. Ambas consideram $a(x_i)$ como a menor distância entre x_i e o objeto mais próximo do mesmo grupo, enquanto $b(x_i)$ é a menor distância entre x_i e um objeto pertencente a outro grupo. SA2 utiliza usa uma função indicadora, em que se $a(x_i) < b(x_i)$, $c(x_i) = 1$. Caso contrário $c(x_i) = 0$. Portanto, em SA1 a silhueta de um objeto é dada pela equação (1), e para a SA2 corresponde à função $c(x_i)$. Por fim, a variante denominada Silhueta Mediana (SM), proposta neste trabalho, é baseada na ST, mas difere por considerar $a(x_i)$ como a distância *mediana* e $b(x_i)$ corresponde à menor distância mediana.

4. Experimentos Computacionais

Para a realização dos experimentos foram utilizadas 51 instâncias bem conhecidas da literatura, com características diversificadas em relação: (i) às dimensões - em objetos (n),

atributos (p) e grupos (k); (ii) estrutura dos dados - grupos bem definidos, coesos e bem separados ou com padrões difusos; (iii) origem dos dados - reais (naturais) ou gerados artificialmente; (iv) problema: para o PA (k é conhecido), Problema de Classificação (classes conhecidas) ou instâncias sem grupos e classes conhecidos. As instâncias possuem entre 61 e 10437 registros (objetos), de 1 a 40 atributos, e de 2 e 12 grupos (quando reportado). Os dados foram normalizados, e todas as instâncias estão disponíveis e relatadas em trabalhos da literatura: conjunto DS2[1], UCI[2], Atlas Brasil[3], SIDRA[4] e DATASUS[5]. Os algoritmos usados, PAM, DBSCAN e BK, estão disponíveis no CRAN[6], nos pacotes *cluster*, *dbscan* e *stats*. Foi utilizado um computador dotado de processador AMD Ryzen 5 de 2.1GHz, com 12 GB de memória e Windows 10.

A Tabela 1 sumariza os resultados de quatro experimentos, e relata: os algoritmos, as distâncias, as instâncias e o percentual de acertos de cada variante do índice silhueta. É considerado um acerto quando o k identificado no experimento estiver a no máximo uma unidade do k considerado ideal.

Exp.	Algoritmo	Dist.	Dataset	ST	SA1	SA2	SS	SM
1	PAM	DE e DM	DS2	93.5%	45.2%	45.2%	96.8%	**100.0%**
2	DBSCAN	DE	DS2	80.6%	58.1%	61.3%	77.4%	**83.9%**
2	DBSCAN	DM	DS2	**61.3%**	48.4%	45.2%	**61.3%**	**61.3%**
3	BK (SM)	DE	DS2	80.6%	35.5%	35.5%	80.6%	**83.9%**
3	BK (ST)	DE	DS2	77.4%	35.5%	35.5%	77.4%	**80.6%**
4	PAM	DE	UCI	37.5%	50.0%	**62.5%**	37.5%	37.5%
4	PAM	DM	UCI	50.0%	50.0%	**62.5%**	50.0%	50.0%

Tabela 1. Percentual de acertos das silhuetas por tipo de distância.

No experimento 1 o algoritmo PAM foi utilizado com o parâmetro k entre 2 e 9 (mínimo e máximo de grupos das instâncias). No experimento 2, para o uso do DBSCAN foi necessário realizar uma calibração de parametros de entrada (DistK). Destaca-se que todos os objetos devem fazer parte das soluções finais, ou seja, objetos inicialmente classificados como ruídos são alocados ao grupo mais próximo. No Experimento 3 optou-se por utilizar somente a DE como métrica, e o (BK) utilizou como critério de seleção do grupo a ser particionado todas as versões da silhueta abordadas nesse trabalho.

No Experimento 4 o algoritmo PAM foi aplicado em 8 instâncias de classificação do repositório do UCI (n entre 106 e 10437, p entre 7 e 40 e k (classes) entre 2 e 12), com o parâmetro $k \in \{2, 3, ..., 12\}$. Foi utilizada a Estatística de Hopkins (EH) para verificar a existência de tendência de agrupamento nesses instâncias, embora as classes sejam conhecidas. Todos os resultados de EH foram superiores a 0.7, o que indica a existência de uma boa estrutura dos dados [Semaan 2013]. Por fim, o Experimento 5 contemplou as instâncias dos repositórios Atlas Brasil, DATASUS e SIDRA, cujo número ideal de grupos não é conhecido. Novamente a EH foi utilizada, sendo confirmada a tendência à formação de agrupamentos para todas as instâncias. A partir dessa tabela, observa-se, nos Experimentos 1, 2 e 3, melhor performance (acertos) da SM frente às demais alternativas e no Experimento 4, ocorre uma inversão, principalmente considerando a SA2 e a DE;

[1]Conjunto de Dados DS2 utilizado em diversos trabalhos e reportados em [Semaan 2013].

[2]*University of California, Irvine - Machine Learning Repository* (https://archive.ics.uci.edu/ml).

[3]Atlas do Desenvolvimento Humano no Brasil (https://dados.gov.br/dataset/atlasbrasil).

[4]Sistema IBGE de Recuperação Automática (https://sidra.ibge.gov.br/acervo).

[5]Ministério da Sáude - OpenDataSUS (https://opendatasus.saude.gov.br).

[6]*The Comprehensive R Archive Network* (https://cran.r-project.org).

fato que pode ser decorrente do grau de assimetria dos dados, o que tende a produzir valores $b(x_i)$ bem superiores aos de $a(x_i)$, implicando alta prevalência de $c(x_i) = 1$. Nesse último experimento, após obter as soluções via algoritmo PAM, a coincidência dos valores ideais de k entre os pares de silhuetas foi calculada, bem como os k associados aos maiores valores de cada versão, conforme apresenta a Tabela 2.

Distância	ST e SA1	ST e SM	SA1 e SA2	SA1 e SM	SS e SM
Euclidiana	**83,3%**	**83,3%**	75,0%	**83,3%**	**83,3%**
Manhattan	75,0%	75,0%	**100,0%**	66,7%	**91,7%**

Tabela 2. Experimento 5 - coincidência de k por tipo de distância.

5. Conclusões e Trabalhos Futuros

Neste trabalho foram estudadas e avaliadas 4 variantes (3 da literatura) para o índice de validação silhueta, medida utilizada para verificar a qualidade de um agrupamento e que combina coesão e separação. A SM, baseada na distância, foi aplicada com o intuito de tornar o índice menos suscetível a valores extremos. Para avaliar o desempenho dessas variantes foram realizados 5 experimentos que contemplaram os algoritmos PAM, DBSCAN e BK, 51 bases de dados da literatura e dois tipos de distância. Nos experimentos 1 e 2, a combinação da SM propiciou um percentual de acertos superior a 80%. No experimento 3 o destaque ocorreu com o uso da SM tanto como critério de seleção quanto na avaliação das soluções. Já nos experimentos 4 e 5, a DM ocasionou maiores percentuais de coincidência entre os pares de silhuetas que a DE, em geral. Em especial, no experimento 5 a correspondência do valor de k ideal foi de 100% entre as SA1 e SA2.

Com base nos resultados obtidos, a SM pode constituir-se como uma boa opção, quando comparada à Silhueta Tradicional, nos casos em que as bases de dados possuem grupos bem estruturados. Como trabalhos futuros pretende-se analisar com mais profundidade as vantagens e desvantagens de cada uma das versões do índice. Além disso, novos experimentos serão realizados com outros algoritmos, como o *Clustering Large Applications* (CLARA) e algoritmos hierárquicos como *Single Linkage* e *Complete Linkage*.

Referências

Amorim, E. R. (2013). *Novos Índices Relativos para a Identificação da Quantidade Ideal de Grupos*. Trabalho de conclusão de curso, Universidade Anhanguera, Niterói - RJ.

Bussab, W. O., Miazaki, E. S., and Andrade, D. F. (1990). *Introdução à Análise de Agrupamentos*. IME - USP, São Paulo.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Hruschka, E. R., Campello, R. J. G. B., and Castro, L. N. (2004). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. In *IEEE International Conference on Data Mining*, pages 403–406.

Kaufman, L. and Rousseeuw, P. J. (1989). *Finding Groups in Data - An Introduction to Clusters Analysis*. Wiley-Interscience Publication.

Semaan, G. S. (2013). *Algoritmos para o Problema de Agrupamento Automático*. Tese de doutorado, Universidade Federal Fluminense, Niterói - RJ.