ANÁLISE DE AGRUPAMENTOS APLICADA AO ESTUDO DE INSTITUIÇÕES DE ENSINO SUPERIOR PÚBLICAS

Rodrigo Barbosa de Oliveira[1], José André de Moura Brito[2]

Introdução

A análise de agrupamentos é uma técnica de análise multivariada que agrega um conjunto de métodos que são aplicados para determinação de grupos a partir de um conjunto de objetos definidos por certas variáveis. Em particular, neste trabalho, tal técnica foi aplicada em uma base de dados constituída por 286 instituições de ensino superior (IES) públicas brasileiras. O objetivo foi definir grupos homogêneos de IES, tomando por base um razoável conjunto de variáveis associadas às instituições, seus docentes e discentes (dados do INEP).

De forma a produzir soluções de boa qualidade, isto é, que propiciem uma boa segmentação das IES, em uma das etapas do processo de agrupamento, foi aplicado um algoritmo heurístico baseado na metaheurística algoritmos genéticos de chaves aleatórias viciadas [Gonçalves and Resende, 2011]. Tal algoritmo foi aplicado para selecionar, a partir do conjunto original de variáveis, um subconjunto de variáveis que, ao ser utilizado como um dos parâmetros de entrada de dois algoritmos de agrupamento não hierárquico, produzissem agrupamentos mais homogêneos. Em uma fase posterior, as soluções associadas aos agrupamentos foram avaliadas, mediante a aplicação dos índices de silhueta [Kaufman and Rousseeuw, 1990] e de Calinski-Harabasz [Calinski and Harabasz, 1974]. Também foram realizados experimentos prévios, por meio da utilização da Estatística de Hopkins [Banerjee, 2004], com o objetivo de identificar se a base de dados composta pelas 286 instituições (e as 75 variáveis) tinha tendência à formação de agrupamentos.

Objetivos

Este trabalho tem como objetivo principal aplicar dois algoritmos de agrupamento e índices de validação para definir grupos de IES públicas brasileiras. Ainda neste sentido,

[1]Escola Nacional de Ciências Estatísticas (ENCE), rodrigobarbosa.deoli@gmail.com

[2]Escola Nacional de Ciências Estatísticas (ENCE), jambrito@gmail.com

face ao substancial número de variáveis (mais de 70), também foi aplicado um algoritmo para selecionar um subconjunto de variáveis, para produzir grupos mais homogêneos.

Material e Método

Toda manipulação, análise e visualização gráfica de dados, presentes nesse trabalho, foram realizadas utilizando a linguagem R [R Core Team, 2018]. Os algoritmos de agrupamento estão disponíveis em funções dos pacotes base e 'cluster' [Maechler et al., 2018]. Os índices de validação foram calculados a partir da aplicação de funções do pacote 'clusterSim' [Walesiak, 2018]. A manipulação de dados foi auxiliada pela coleção de pacotes 'tidyverse' [Wickham, 2017], também utilizada para visualização gráfica de dados, assim como o pacote 'plotly' [Sievert, 2018].

Para formação dos grupos, foram utilizados dados de docentes e discentes (variáveis) do Censo da Educação Superior (ano de 2017) disponíveis no sitedo INEP, relativos a 286 instituições de ensino superior (IES). Após uma pré-seleção de variáveis consideradas, a padronização das mesmas e exclusão de variáveis que apresentavam valores extremos, foram selecionadas 75 variáveis quantitativas. Entre elas, temos: proporção de bolsas de monitoria e de iniciação científica, quantidade de alunos, docentes, proporção de alunos oriundos de escolas públicas etc.

Conforme comentado na introdução, os métodos disponíveis em análise de agrupamentos possibilitam resolver o seguinte o problema: dado um conjunto X constituído por n objetos ($X=\{x_1, x_2,..., x_n\}$), tal que cada objeto x_i tem f atributos, isto é, $x_i=(x_{i1}, ..., x_{if})$, e definida uma métrica que permite avaliar o grau de dissimilaridade entre os objetos, por exemplo, a distância euclidiana, busca-se alocar os n objetos em k grupos, de forma que os grupos tenham alto grau de homogeneidade internamente e baixo grau de homogeneidade entre si. Tal homogeneidade depende da métrica e de uma função objetivo definidas a priori para o problema de agrupamento, sendo tais grupos definidos, mediante a aplicação de algoritmos não hierárquicos e/ou hierárquicos. No presente trabalho, foram considerados dois algoritmos de agrupamento não-hierárquico: k-medoids, utilizando as distâncias euclidiana, de Manhattan e de Mahalanobis, e o k-means. A eficácia (qualidade dos grupos) desses algoritmos foi avaliada utilizando o índice de silhueta [Kaufman and Rousseeuw, 1990], que também serviu de base para determinar o número ideal de grupos, assim como o índice de Calinski-Harabasz [Calinski and Harabasz, 1974].

O algoritmo k-means também foi utilizado em um dos procedimentos de um algoritmo baseado no algoritmo genético de chaves aleatórias viciadas [Gonçalves and Resende,

2011] (Biased Random-Key Genetic Algorithm), que consiste em uma metaheurística evolutiva para problemas de otimização. O algoritmo heurístico foi utilizado para determinar o melhor subconjunto de variáveis, à luz do índice de silhueta, calculado após aplicação do algoritmo k-means.

Resultados e Discussão

Considerando, inicialmente, as 75 variáveis, foi aplicada sobre a base de 286 IES a Estatística de Hopkins [Banerjee, 2004] e, em seguida, foram aplicados os algoritmos k-means e k-medoids (k variando entre 2 e 6). No que concerne à Estatística de Hopkins, foi obtido o valor de 0.8347547, indicando que a base de dados tem tendência para formação de grupos (essa estatística varia entre 0 e 1).

Considerando as 75 variáveis, independente do algoritmo e do número de grupos, foram encontrados valores abaixo 0,14 em relação ao índice de silhueta (silhueta média), o que, pela literatura, indica que não foi encontrada uma estrutura razoável quanto aos grupos formados.

Uma primeira alternativa simples para tentar melhorar o valor da silhueta e, consequentemente, a qualidade das soluções (grupos produzidos pelos algoritmos), foi realizar um experimento computacional, onde foram selecionadas 500 amostras aleatórias simples, estando cada amostra associada a um subconjunto de variáveis provenientes das 75 variáveis.

A Figura 1 traz boxplots associados à distribuição dos valores do índice de silhueta obtidos para as soluções associadas com número de grupos entre 2 e 6, considerando aplicação dos algoritmos k-means e k-medoids em cada uma das amostras de variáveis.

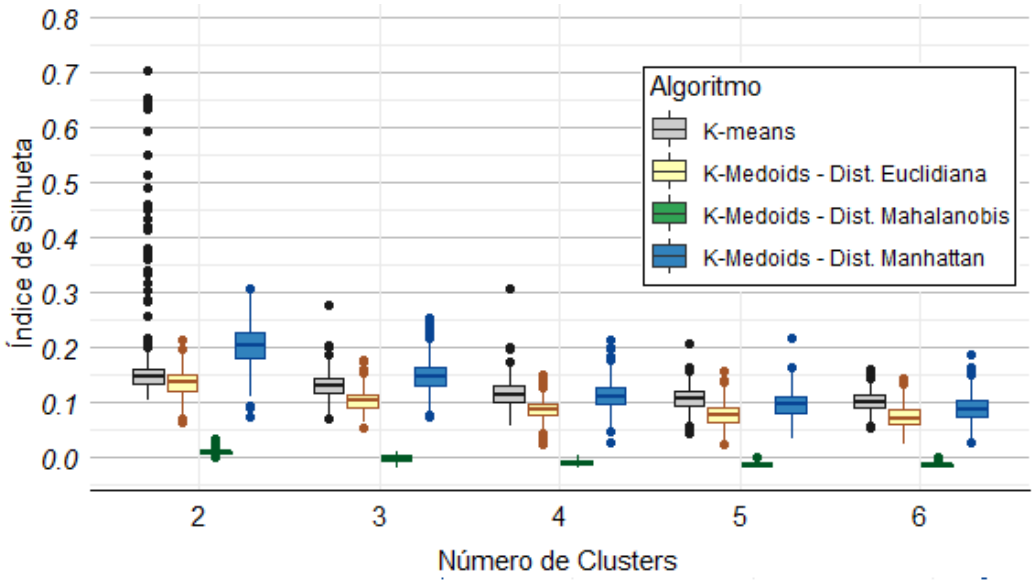

Figura 1 - Desempenho por algoritmo de agrupamento

Analisando a Figura 1, observa-se que o algoritmo k-medoids, com distância de Manhattan, e o algoritmo k-means apresentaram os maiores valores para o índice de silhueta para número de grupos igual 2 (k=2). Todavia, o valor mediano do índice de silhueta ficou da ordem de 0.2, o que indica uma estrutura de agrupamento fraca.

Face aos baixos valores observados para o índice de silhueta ao adotar a alternativa 1, buscou-se, como segunda alternativa, a aplicação de um método de otimização. Neste sentido, foi implementado um algoritmo baseado na metaheurística BRKGA, denotado por BRKGAVAR. Em linhas gerais, o algoritmo BRKGAVAR produz, durante as suas q iterações, várias soluções, sendo cada solução correspondente a um subconjunto de variáveis selecionadas a partir das 75 variáveis da base. Para cada solução produzida, aplica-se o algoritmo k-means e avalia-se o valor da silhueta. A idéia do algoritmo é produzir, em cada iteração, soluções cada vez melhores, isto é, subconjuntos de variáveis que, ao serem utilizadas em todas as IES, produzirão silhuetas maiores. Essas soluções são melhoradas a partir da aplicação de procedimentos específicos do BRKGA, quais sejam: cruzamento e mutação. A seguir, na Tabela 1, são apresentados os resultados obtidos a partir da aplicação do BRKGAVAR, considerando o melhor valor de silhueta encontrado para cada valor de k, após a q gerações do desse algoritmo.

Tabela1 – Índice de silhueta para variáveis selecionadas pelo BRKGAVAR

Número de Grupos	2	3	4	5	6	7	8	9	10
Índice de Silhueta	0,82	0,60	0,44	0,37	0,33	0,31	0,34	0,36	0,35

Os resultados obtidos a partir da aplicação do BRKGAVAR indicam uma melhora razoável no índice. Todavia, independente da estratégia adotada (seleção de amostras de variáveis ou aplicação do BRKGAVAR), as melhores soluções estão associadas aos números de grupos iguais a 2 ou 3.

Conclusões

Com o objetivo de agrupar as instituições de ensino superior públicas, o presente trabalho trouxe uma proposta de aplicação de algoritmos de agrupamento e de seleção de variáveis que produzissem agrupamentos de boa qualidade à luz do índice de Silhueta,

utilizado como o critério principal para avaliar as soluções produzidas pelos algoritmos de agrupamento.

Os próximos desdobramentos desse trabalho contemplarão a aplicação dos algoritmos de agrupamento k-means, k-medoids e DBSCAN [Ester et al. 1996], considerando a base de dados segmentada pelas cinco grandes regiões do Brasil e por categorias administrativas (Federal, Estadual e Municipal). E, por fim, uma análise crítica dos agrupamentos formados pelas IES.

Referências

[Banerjee, 2004] Banerjee, A. (2004). Validating clusters using the hopkins statistic. IEEE International Conference on Date of Conference, 1:149-153.

[Calinski and Harabasz, 1974] Calinski, R.; J. Harabasz (1974).A dendrite method for cluster analysis. Communications in Statistics 3, 1–27.

[Ester et al. 1996] 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 Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, page 226–231.

[Gonçalves and Resende, 2011] Gonçalves, J.F.; Resende, M. G. C. (2011). Biased random-key genetic algorithms for combinatorial optimization. J. of Heuristics, 17:487-525.

[Kaufman and Rousseeuw, 1990] Kaufman, L.; Rousseeuw, P. J. (1990). An introduction to cluster analysis. John Wiley and Sons.

[Maechler et al., 2018] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2018). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1.

[R Core Team, 2018] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.

[Sievert, 2018] Sievert, C (2018). plotly for R. URL https://plotly-book.cpsievert.me.

[Walesiak, 2018] Walesiak, M.; Dudek,A. (2018). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package version 0.47-2. URL https://CRAN.R-project.org/package=clusterSim.

[Wickham, 2017] Wickham, H.(2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. URL https://CRAN.R-project.org/package=tidyverse.