Implementação dos métodos computacionais do artigo "Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets", Maitra et al. (2012)*

Alexandre Gandini

Setembro de 2020.

Resumo

Apresentamos a implementação dos métodos computacionais propostos por Maitra, Melnykov e Lahiri (2012), assim como uma breve introdução, revisão da metodologia, simulação de Monte Carlo e aplicação em dados reais. **Palavras-chave**: Estatística Computacional. Bootstrap. Agrupamentos.

1 Introdução

O problema da análise de agrupamentos consiste em particionar *datasets* de forma que observações similares entre si pertençam a um mesmo grupo e, ao mesmo tempo, para este grupo específico, que os pontos apresentem características diversas em relação às observações pertencentes aos demais.

Na literatura de métodos de agrupamentos, há basicamente duas abordagens para tal problema:

- Paramétrica ou *model-based*: cada observação pertence a um componente de uma mistura de distribuições;
- Não paramétrica, ou *model-free*: métodos baseados em distâncias.

Uma das maiores dificuldades em se trabalhar com análise de agrupamentos é, nas duas abordagens, a definição de K, a quantidade de grupos.

O artigo em questão propõe um método para estimar o melhor número K para a abordagem *model-free*, utilizando como critério a função objetivo melhora da soma dos quadrados das distâncias dentro dos grupos:

*Trabalho submetido como requisito parcial para a conclusão da disciplina de Estatística Computacional do Mestrado em Estatística do PPGEst UFRGS.

$$S_{K;K^*} = W_K - W_{K^*}$$

Onde

$$W_K = \sum_{i=1}^{n} \sum_{k=1}^{K} I_{i \in k}(\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'(\boldsymbol{x}_i - \boldsymbol{\mu}_k)$$

E μ_k é a média dos pontos de cada grupo, k um grupo específico, n o tamanho da amostra e x uma observação. $I_{i \in k}(x)$ é a função indicadora da i-ésima observação pertencer ao grupo k.

Para os autores, um $K^* > K$ implica em um modelo mais complexo (em termos de quantidade de grupos), portanto o K^* estimado deve ser aquele no qual mostre-se mais ajustado aos dados, em comparação aos outros K, ou seja, aquele no qual a melhora no ajuste não se deu puramente em razão da aleatoriedade.

Note que $S_{K;K^*}$ é positiva, visto que W_K sempre diminui a medida em que a quantidade de grupos aumenta. O objetivo é testar se a melhora em W_K é significativa ao ajustar K^* grupos contra uma hipótese nula de K grupos.

No trabalho de Maitra, Melnykov e Lahiri (2012), o que se propõe é uma abordagem baseada em bootstrap, livre de premissas paramétricas, para quantificar a significância da quantidade de grupos em um dataset. O que é testado é, dada uma hipótese nula de K grupos, se uma alternativa mais complexa K^* leva a um melhor ajuste dos dados, reduzindo significativamente a soma dos quadrados das distâncias dentro dos grupos.

Outras iniciativas para testar a significância da quantidade de grupos incluem Liu et al. (2008) e Kimes et al. (2017), em especial no contexto de alta dimensionalidade e baixa quantidade amostral (HDLSS), usando premissas de modelos paramétricos de misturas.

Desde a sua publicação, o trabalho já foi citado em 31 outros artigos[1], inclusive por Cybis e Valk (2018) e Valk e Cybis (2020).

2 Metodologia

O aspecto central da metodologia proposta é o cálculo do valor-p da estatística de teste em termos da probabilidade de um $S_{K;K^*}$ obtido de K-grupos ser maior do que um $S_{K;K^*}$ calculado a partir dos dados.

Na seção *2.3 Obtaining a Reference Distribution*, os autores propõem os seguintes passos para o algoritmo:

- Para cada $i = 1, 2, ..., n$ elemento da amostra, gera aleatóriamente um vetor unitário p-dimensional, utilizando uma distribuição Normal padrão multivariada, da seguinte forma: $W_i = \frac{Z_i}{||Z_i||}$, onde Z_i é o vetor gerado aleatoriamente, $||Z_i||$ é a sua norma, resultando no vetor W_i com comprimento unitário.
- Em seguida, gera o resíduo simulado $e_i^* = ||\hat{e}_{li}^*||W_i$ onde $\hat{e}_{li}^*$ é obtido através de uma permutação aleatória dos elementos da amostra.

[1] Consulta em via Google Scholar em agosto de 2020.

- Assim, o conjunto dos resíduos reamostrados e_i^* mantém as normas do conjunto e_i, mas com direções diferentes, seguinto o effeito de W_i.
- Adicionando os $\hat{e}_i^*$ às médias dos grupos $\hat{\mu}_k$, após tê-las reescaladas por $\hat{\sigma}$, resulta em uma realização reamostrada do dataset sob H_0 (a hipótese de que a quantidade verdadeira de clusters é k), da seguinte forma:

$$\boldsymbol{X}_i^* = \sum_{k=1}^{K} I_{i\in k}(\boldsymbol{x}_i)\boldsymbol{\mu}_k + \hat{\sigma}\boldsymbol{e}_i^*$$

Onde:

$$\hat{\sigma} = \frac{1}{np}\sum_{i=1}^{n}\sum_{k=1}^{K} I_{i\in k}(\boldsymbol{x}_i)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{k}})'(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{k}})$$

E $\hat{\boldsymbol{\mu}}$ são os centros estimados pelo algoritmo de clusterização, no caso do presente trabalho foi utilizado o *K-Means*, na implementação padrão da biblioteca *Scikit-Learn*.

A realização reamostrada dos dados sob H_0 é chamada de:

$$\Xi^* = \{X_1^*, X_2^*, ..., X_n^*\}$$

- Repete-se o procedimento M vezes para obter realizações reamostradas dos dados $\Xi_1^*, \Xi_2^*, ..., \Xi_M^*$. Para cada Ξ_j^*, calcula-se a estatística de teste $s_{j,(K;K^*)}^*$ para $j = 1, 2, ..., M$.
- O valor-p de s_{K,K^*} é então estimado através da proporção de vezes em que a estatística dos dados reamostrados $s_{j,(K;K^*)}^*$ excedeu a estatística dos dados originais. Formalmente:

$$\text{valor-p} = \frac{1}{M}\sum_{j=1}^{M} I(s_{j,(K;K^*)}^* > s_{K,K^*})$$

3 Implementação

Tivemos acesso à implementação do artigo original pelos autores na linguagem de programação C, fornecida pelo professor Márcio Valk. Entretanto, ao analisar o código, ele se mostrou extremamente confuso, de forma que não utilizamos qualquer parte daquele trabalho, e implementamos uma versão completamente nova em Python.

O código pode ser acessado em <https://github.com/alexgand/computacional_ppgest_2020>.

4 Estudo de Simulação e Aplicações com Dados Reais

4.1 Dados Reais

Iniciamos com a análise de dados reais, através do *Wine Dataset*[2], que contém 178 observações de tipos diferentes de vinhos, cujas classes foram anotadas por humanos em três categorias (classes 0, 1 e 2), com as seguintes variáveis, todas numéricas:

[2] <https://scikit-learn.org/stable/datasets/index.html>

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

Como as variáveis estão em ordens de grandeza diferentes, para que o algoritmo de clusterização *K-Means* desse igual importância para cada uma delas, as variáveis foram padronizadas para resultarem em novas variáveis com média igual a zero e desvio padrão unitário.

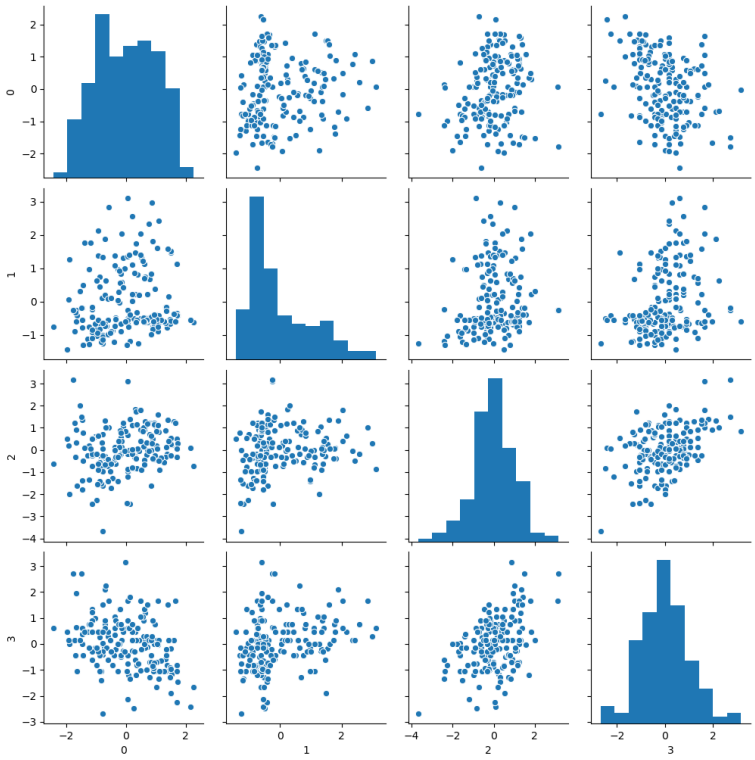
A Figura 1 traz a visualização dos histogramas e gráficos de dispersão par-a-par para as quatro primeiras variáveis apenas, para não poluir demais a imagem.

Foram testados as seguintes possíveis quantidades verdadeiras de grupos: 1, 2, 3, 4, 5. Sabemos que os dados possuem uma quantidade verdadeira de 3 grupos, o que condiz com a saída do procedimento de Maitra, Melnykov e Lahiri (2012), conforme tabela abaixo:

Quantidade de grupos	Valor-p
1	0.00
2	0.02
3	0.08
4	0.08
5	0.08

Utilizando o nível de significância de 5%, segundo os autores, o algoritmo conseguiu capturar a informação de que um $k = 3$ produz um resultado ótimo, no sentido em que modelos mais complexos, digamos com $k = 4, 5, \ldots$ não passam no teste de significância da estatística $S_{K,K*}$.

Figura 1 – Histograma e dispersão par-a-par para as primeiras variáveis do *Wine Dataset*



4.2 Simulação de Monte Carlo

Em seguida, foram geradas 1000 amostras controladas de dados bivariados, com tamanho 200 e dois centros de grupos, através da função `<make_blobs()>`[3] do pacote Scikit-Learn[4].

A Figura 2 traz a visualização da primeira amostra bivariada.

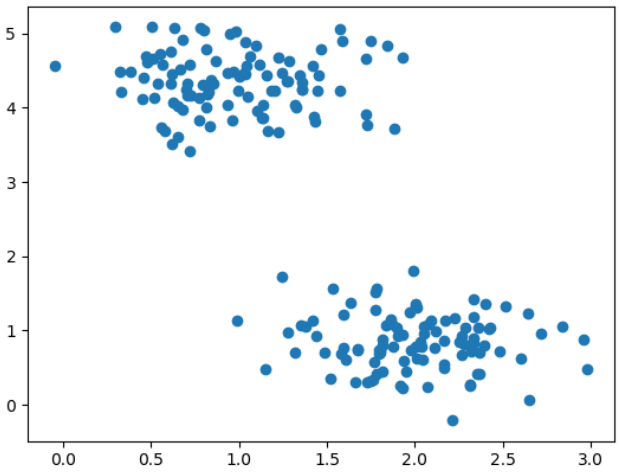
Segue abaixo tabela dos valores-p para a primeira replicação, indicando a quantidade correta de dois grupos:

Quantidade de grupos	Valor-p
1	0.00
2	0.08
3	0.08
4	0.09
5	0.10

Após as 1000 replicações, o algoritmo proposto pelos autores foi capaz de encontrar a verdadeira quantidade de grupos (dois) em 100% dos casos.

[3] Generate isotropic Gaussian blobs for clustering.
[4] <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_blobs.html>

Figura 2 – Dispersão para uma amostra bivariada simulada



5 Conclusões

Apresentamos a implementação dos métodos computacionais propostos por Maitra, Melnykov e Lahiri (2012) para a descoberta do melhor número k de grupos em um problema de agrupamento para dados multivariados.

O algoritmo aqui apresentado se aplica somente à casos onde os grupos possuem forma esférica, sendo utilizado o clusterizador *K-Means*, entretanto o artigo também abrange grupos com geometrias mais genéricas, elipsoidais, utilizando algoritmos de clusterização hierárquica para este caso.

No caso do presente trabalho, o resultado foi satisfatório no sentido de encontrar a quantidade verdadeira de grupos, tanto trabalhando com dados reais (*Wine Dataset*), quanto com dados simulados.

Pensando em estudos futuros, é possível ampliar o trabalho dos autores para englobar também medidas de distância não euclidiana ou similares quando da clusterização: útil para casos de agrupamento de texto, em que comumente a distância de cosseno (ângulo entre vetores de duas observações) é a que apresenta melhor resultado prático.

Outros possíveis desenvolvimentos são: adaptar o trabalho para dados discretos e possibilitar que os grupos tenham geometria ainda mais genérica do que formas elipsoidais.

6 Repositório

O código produzido para realizar o presente trabalho está disponível em <https://github.com/alexgand/computacional_ppgest_2020>.

Referências

CYBIS, G.; VALK, M. Clustering and classification problems in genetics through u-statistics. *Journal of Statistical Computation and Simulation*, v. 88, p. 1882–1902, 2018. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/00949655.2017.1374387>. Citado na página 2.

KIMES, P. K. et al. Statistical significance for hierarchical clustering. *Biometrics*, p. 811–821, 2017. Citado na página 2.

LIU, Y. et al. Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, v. 103, p. 1281–1293, 2008. Citado na página 2.

MAITRA, R.; MELNYKOV, V.; LAHIRI, S. N. Bootstrapping for significance of compact clusters in multidimensional datasets. *Journal of the American Statistical Association*, v. 107, p. 378–392, 2012. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.2011.646935>. Citado 4 vezes nas páginas 1, 2, 4 e 6.

VALK, M.; CYBIS, G. U-statistical inference for hierarchical clustering. *Journal of Computational and Graphical Statistics*, p. 1–11, 2020. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.2011.646935>. Citado na página 2.