Seleção de Atributos para Analise de Agrupamento de Dados

Um Estudo Comparativo

Sibelius Seraphini
Instituto de Ciências Matemáticas e de Computação (ICMC) - Universidade de São Paulo (USP)
Avenida Trabalhador São-carlense, 400 - Centro São Carlos, Brasil
sibelius@usp.br

ABSTRACT

O objetivo da seleção de atributos para aprendizado não-supervisionado é encontrar o menor subconjunto de atributos que melhor revela agrupamentos naturais interessantes. O problema se torna ainda mais complexo se o número de agrupamento não é conhecido. Consequentemente, existe a necessidade de encontrar o número de grupos concomitantemente com o subconjunto de atributos, que geralmente são inter-relacionados. Na literatura existem alguns trabalhos que propõem algoritmos para seleção de atributos para agrupamento de dados. Desse modo, este artigo visa realizar um estudo comparativo entre alguns dos algoritmos propostos na literatura, com o objetivo de identificar qual deles é melhor para um determinado cenário. Os algoritmos foram aplicados sistematicamente em conjuntos de dados gerados segundo um design fatorial experimental. Dois critérios de validação externa foram utilizado para o realizar o comparativo entre os algoritmos. Três aspectos foram levados em conta para avaliar os algoritmos: qualidade do agrupamento obtido (critério externo), estimativa do número de agrupamentos, e quantidade de atributos selecionados. Os algoritmos *clustvarsel* e *vscc* apresentaram os melhores resultados, não apresentando diferença estatisticamente significativas entre si. A ampliação do conjunto de dados sinteticos e a acréscimo de outros conjuntos de dados reais podem constituir um benchmark para novos algoritmos de seleção de atributos para agrupamento de dados.

1. INTRODUÇÃO

O agrupamento de dados é um problema fundamental na mineração de dados, em que se visa determinar um conjunto finito de categorias para descrever um conjunto de dados de acordo com as similaridades entre os seus objetos [17, 19, 7, 10] Nesse problema, a estrutura de maior interesse para o investigador pode ser melhor representada usando apenas alguns dos atributos. Além disso, utilizando-se somente os atributos relevantes do conjunto de dados pode levar a um melhor modelo de agrupamento de dados para descrever os dados futuros, e também menos atributos podem produzir uma melhor partição de dados em grupos mais próximos da estrutura dos grupos verdadeiros. Nesse sentido, uma tarefa importante de analise estatística e mineração de dados é a seleção de atributos. A seleção de atributos visa escolher um subconjunto dos atributos originais, eliminando os redundantes, não informativos, e ruidosos. Existem vários potencias benefícios da seleção de atributos como, por exemplo [11]: facilitar o entendimento e visualização dos dados, reduzir as necessidades de medição e armazenamento, reduzir o tempo de treinamento e utilização e diminuir a maldição da dimensionalidade para melhorar o desempenho de predição.

A literatura para seleção de atributos para classificação (aprendizado supervisionado) é vasta [3, 9, 1, 20]. Para o aprendizado supervisionado, os algoritmos de seleção de atributos maximizam alguma função de capacidade preditiva. Porque são dados as classes dos objetos, é natural manter somente os atributos que são relacionados com essas classes. Mas no caso do aprendizado não-supervisionado, não se conhece a priori as classes (grupos) dos objetos. Além disso, o problema se torna ainda mais complexo se o número de grupos da base de dados não é conhecido a priori. Consequentemente, existe a necessidade de encontrar o número de grupos concomitantemente com o subconjunto de atributos, que geralmente são inter-relacionados.

Levando isto em consideração, este artigo visa realizar um estudo comparativo de alguns algoritmos de seleção de atributos para agrupamento de dados. Esse estudo tem dois objetivos: o primeiro é verificar se a seleção de atributos melhora o modelo de agrupamento de dados; o segundo é identificar qual dos algoritmos investigados melhor consegue identificar atributos relevantes, melhorando assim o agrupamento de dados. Desse modo, dois algoritmos de agrupamentos serão treinados utilizando todos os atributos, k-means [13, 22] e EM ou baseado em modelo [8, 5, 14]. Ademais, quatro algoritmos de seleção de atributos: sparse K-means [29, 22], clustvarsel [26], vscc [2], baseado na silhueta simplificada (SS) [15]; serão utilizados para selecionar os atributos relevantes, após os algoritmos k-means e EM serão treinados utilizando o subconjunto de atributos selecionados. A comparação entre o grupo que considera todos os atributos e o que considera os atributos selecionados, assim como a comparação entre os algoritmos de seleção de atributos investigados será feita utilizando dois critérios de validação externa: índice de Rand ajustado [16] e Jaccard [17].

O restante desse artigo é estruturado como segue. A Seção 2 apresenta o conceito da seleção de atributos e os algoritmos de seleção de atributos para agrupamento de dados que serão investigados nesse estudo comparativo. Além disso, também descreve sobre critérios de validação de agrupamento e em particular sobre os dois critérios que serão empregadas nesse estudo comparativo. Na Seção 3 é descrito os conjuntos de dados utilizados, assim como a metodologia empregada e também apresenta-se os resultados e discussões. Finalmente, a Seção 4 aborda as conclusões desse estudo comparativo.

2. TRABALHOS RELACIONADOS

2.1 Seleção de Atributos para Agrupamento de Dados

O objetivo da seleção de atributos para aprendizado não-supervisionado é encontrar o menor subconjunto de atributos que melhor revela agrupamentos naturais interessantes (grupos) [6]. Os métodos de seleção de atributos são classificados em três tipos:

- *abordagem de filtro*: atributos são avaliados individualmente usando o objetivo da tarefa e um valor de relevância é calculado que é utilizado para ordenar os atributos
- *abordagem wrapper*: atributos são avaliados num subconjunto, um algoritmo que realiza a tarefa de aprendizado é utilizado para essa avaliação
- *abordagem embutida*: a seleção dos atributos mais importantes é parte integrada do tarefa de aprendizado

Um algoritmo que utilizado a abordagem *wrapper* é o proposto por [15]. [15] apresenta um algoritmo que não necessita do número de grupos a priori, encontrando tanto o número de grupos e o subconjunto de atributos que exibe o maior valor de silhueta simplificada [15], critério de validação interna. O algoritmo emprega a busca sequencial *forward search* [21], que começa com um subconjunto vazio e sucessivamente adiciona atributos; o algoritmo k-means em conjunto com o critério silhueta simplificada é utilizado para avaliar o subconjunto de atributos.

Em contrapartida, [29] modela o problema de seleção de atributos como um problema de otimização, no qual cada atributo possui um peso. Nessa abordagem, os atributos selecionados são aqueles que apresentam peso diferente de zero, após o problema ser maximizado.

Outra técnica de seleção de atributos é dada por [26], onde múltiplos modelos da família MCLUST são comparados usando fatores aproximados de Bayes [18]. Essa técnica encontra-se disponível no pacote *clustvarsel* [4] do *R*. Contudo, por causa do número de parâmetros que exige estimativa de alguns modelos *mclust* é quadrático na dimensionalidade dos dados, o pacote *clustvarsel* pode ser muito lento em grandes dimensões. Além disso, essa técnica às vezes pode levar a resultados inferiores quando comparados com o uso de *mclust* sozinho [23]

Além disso, [2] introduz uma técnica de seleção de atributos que é intuitiva e computacionalmente eficiente, podendo ser utilizada tanto para problemas de agrupamento de dados como classificação. O método proposto consiste em definir uma relação entre a variância entre grupos e a correlação entre as variáveis. Cinco relações foram escolhidas para selecionar os subconjuntos de atributos. Os subconjuntos são avaliados utilizando o critério de informação Bayesiana (BIC) [27], que mede o grau de incerteza do modelo.

2.2 Critérios de Validação Externos

Índices ou Critérios de validação de agrupamento [12, 24] são usados para avaliar a validade das partições obtidas por um algoritmo de agrupamento pela comparação de suas características com uma referencia ou partição ideal. Esses índices calculam um valor relativo ou absoluto para a semelhança entre as diferentes partições dos mesmos dados.

Esses índices podem ser classificados em critérios externos, internos e relativos. Os primeiros assumem que a avaliação da partição é baseada em uma estrutura pré-especificada que reflete a estrutura natural do conjunto de dados. Para os segundos, a avaliação é realizada em termos de quantidades obtidas a partir dos valores dos atributos que descrevem os grupos, que medem quão separados e compactos são. Os últimos assumem que é realizado uma comparação entre uma partição com outras obtidas utilizando diferentes parâmetros do algoritmo de agrupamento de dados ou outro algoritmo de agrupamento. Geralmente são critérios internos capazes de quantificar a qualidade de agrupamentos.

A ideia principal dos critérios externos é que quando dois exemplos pertencem a um grupo natural nos dados, eles usualmente aparecem no mesmo grupo quando diferentes partições são obtidas. Isso significa que duas partições que mantêm essas coassociações são mais similares e representam uma estrutura similar para o conjunto de dados. Para calcular esses índices, a coincidência de cada par de exemplos nos grupos de dois agrupamentos tem de ser contado, existem quatro possíveis casos:

- a: Número de pares de objetos de dados que pertencem a mesma classe em C e ao mesmo grupo em P.
- b: Número de pares de objetos de dados que pertencem a mesma classe em C e a diferente grupos em P.
- c: Número de pares de objetos de dados que pertencem a diferentes classes em C e ao mesmo grupo em P.
- d: Número de pares de objetos de dados que pertencem a diferentes classes em C e a diferente grupos em P.

Destes valores diferentes índices de similaridade podem ser definidos para comparação de duas partições:

- *Adjusted Rand Index (ARI)* [16, 17]:

$$AR = \frac{a - \frac{(a+c)(a+b)}{a+b+c+d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(b+c)}{a+b+c+d}}$$

- *Jaccard Coeffient* [17, 19]:

$$J = \frac{a}{a+b+c}$$

Os dois critérios de validação externos apresentados serão utilizados na metodologia proposta para comparar os algoritmos de seleção de atributos. A próxima Seção descreve a metodologia para realizar este estudo comparativo.

3. EXPERIMENTOS

3.1 Conjuntos de Dados

Para comparar os algoritmos de seleção de atributos para agrupamento de dados utilizou-se um conjunto de dados sintéticos. Os conjuntos de dados foram gerados utilizando o pacote *clusterGeneration* que implementa a metodologia descrita em [25]. A Tabela 1 apresenta o fatores de design experimental para gerar os conjuntos de dados. A geração utilizou-se de quatro fatores de design: grupos — define o número de agrupamentos; separação, L — agrupamento próximo, M — agrupamento separado, H — agrupamento bem separado; atributos — atributos relevantes; atributos ruidosos — atributos que não trazer informações sobre o agrupamento. Em resumo, os quatro fatores de design experimental descritos acima foram combinados, assim produzindo um conjunto de 108 (3 número de agrupamentos x 3 graus de separabilidade x 3 números de atributos x 4 números de atributos ruidosos). Seguindo o design de [24] e para obter uma melhor confiança estatística, três replicações independentes foram utilizadas para cada combinação, logo, resultando em 324 conjuntos de dados.

3.2 Metodologia Experimental

Com os conjuntos de dados gerados, dois algoritmos de agrupamentos de dados e quatro algoritmos de seleção de atributos para agrupamento de dados, i.e., k-means [13, 22] e EM [8, 5, 14]; e sparse K-means [29], clustvarsel [26], vscc [2], baseado na silhueta simplificada (SS) [15], foram sistematicamente aplicados para cada conjunto de dados. Os algoritmos k-means e EM utilizam todos os atributos do conjunto de dados. Dentre os algoritmos de seleção de atributos, somente o sparse K-means exige que o número de agrupamento seja fornecido a priori. Nesse sentido, conduziu-se dois experimentos: o primeiro aplica todos os algoritmos acima fornecendo o número de agrupamentos verdadeiros; no segundo aplicou-se somente os algoritmos *clustvarsel* e *vscc*[1], sem que o número de agrupamentos verdadeiros fossem fornecidos. Devido a isso, o primeiro experimento aplica todos os algoritmos acima fornecendo o número de agrupamentos a priori, para que fosse possível o comparativo entre eles.

Para melhorar confiança da analise, os resultados correspondem a dois critérios de validação externos distintos descritos na Seção 2.2, i.e., ARI e coeficiente Jaccard. Histogramas dos resultados sugerem fortemente que as distribuições amostrais observados dificilmente satisfazem a suposição de normalidade. Por essa razão, será adotado o teste não-parametrico de Wilcoxon/Mann-Whitney (W/M-W) A eficiência do teste W/M-W é 0.95 com relação a testes paramétricos como o teste T ou o teste Z mesmo se os dados são normais. Assim, mesmo quando a suposição de normalidade é satisfeito, o teste W/M-W pode ser preferível [28]. Esse teste será usado nesse trabalho para comparar o resultado de cada par de algoritmos como duas amostras.

[1]O algoritmo *SS* não foi empregado por não ser computacionalmene eficient

Ademais, para o segundo experimento, o número de agrupamentos de uma partição eleita como a melhor pelo algoritmo de seleção de atributos é comparado contra o número certo de agrupamentos conhecido para um conjunto de dados particular. O maior número de acertos é 324 (número de conjuntos de dados).

3.3 Resultados Experimental

A Tabela 2 mostra o número de acertos e a porcentagem para o número correto de agrupamentos. Portanto, tanto o algoritmo *clustvarsel* e *vscc* conseguem estimar bem o número correto de agrupamentos (93%) A média de número de acertos dos grupos *clustvarsel* e *vscc* são 0.938 e 0.935; as distribuições nos dois grupos não difere significativamente (Mann–Whitney U = 52650, n1 = n2 = 324, P = 0.436 unicaudal).

Os resultados do primeiro experimento utilizando o critério externo ARI é reportado na Tabela 3. As médias dos valores do ARI — computados sobre toda a coleção de $N_D = 324$ valores disponíveis (um para cada conjunto de dados) — é apresentado na parte inferior da tabela. Os valores mostrados em cada celula do topo da tabela correspondem a diferença entre as médias dos correspondentes pares de algoritmos. Uma célula sombreado indica que a correspondente diferença é estatisticamente significante para nível de $\alpha = 5\%$ (teste unicaudal) com relação ao teste W/M-W.

Vale lembrar que a analise exemplificada na Tabela 3 também foi realizada para o critério externo Jaccard (Apêndice, Tabela 4). Levando em consideração ambas as tabelas, podemos derivar algumas conclusões interessantes. Os algoritmos *clustvarsel*, *EM*, *vscc* apresentam resultados melhores do que os outros algoritmos, porém não apresentam diferenças significativas entre si. O algoritmo baseado na silhueta simplificada apresenta o pior resultado entre os algoritmos.

Ademais, buscou-se verificar se os algoritmos conseguiram eliminar os atributos ruidosos. A Figura 1 apresenta um gráfico boxplot com os valores do tamanho do subconjunto de atributos que foram selecionados por cada algoritmo. Vale destacar, que o boxplot rosa, roxo e marrom correspondem aos valores dos atributos ruidosos, dos atributos não-ruidosos, e do total de atributos, respectivamente. A partir desse gráfico, verificamos que o algoritmo sparseKmeans seleciona praticamente todos os atributos, porém como ele atribui um peso para cada atributo, mesmo atributos com pesos próximo de zero são considerados como selecionados. Por outro lado, o algoritmo baseado em silhueta simplifica seleciona poucos atributos, perdendo informações importantes dos outros atributos para encontrar os agrupamentos. Tanto o *clustvarsel* quanto o *vscc* conseguem eliminar os atributos ruidosos, chegando mais perto do subconjunto contendo somente os atributos relevantes.

Levando em consideração as três analises realizadas nesse estudo comparativo, pode-se derivar que dentre os algoritmos de seleção de atributos investigados os que obtiveram os melhores resultados foram os *clustvarsel* e *vscc*. Ambos conseguem identificar o correto número de agrupamentos (93% de acerto nos conjuntos de dados), também apresentam um resultado superior aos outros algoritmos considerando-se os dois critérios de validação externos utilizados. Além disso,

	Grupos	Separação	Atributos	Atributos Ruidosos	Réplicas	
Fatores	2, 3, 5	L, M, H	2, 5, 10	0, 1, 3, 10	1,2,3	Total
Níveis	3	3	3	4	3	324

Table 1: Característica do design fatorial experimental utilizado para gerar os conjuntos de dados

Algoritmo	Nº de Acertos	%
clustvarsel	304	93.82
vscc	303	93.51

Table 2: Número de acertos de estimativa do número de agrupamentos para cada algoritmo considerando uma coleção de 324 conjuntos de dados[3]

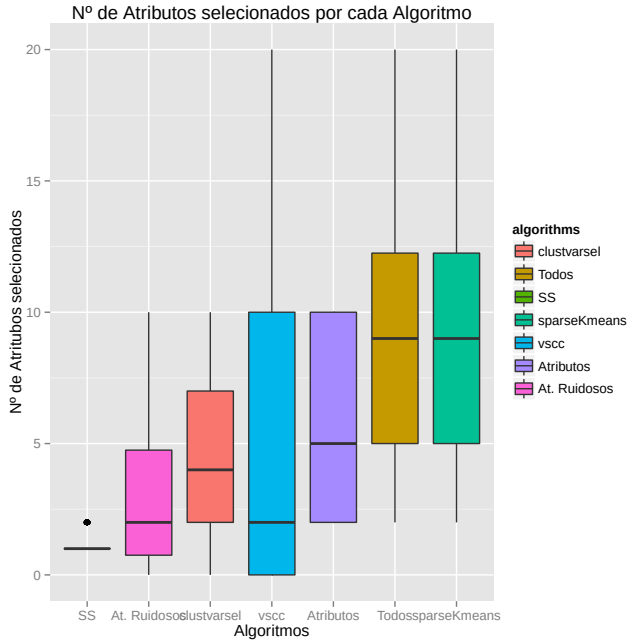

Figure 1: Nº de Atributos selecionados por cada Algoritmo

também conseguem eliminar razoavelmente os atributos irrelevantes.

4. CONCLUSÕES E TRABALHOS FUTUROS

Esse artigo apresentou um estudo comparativo entre algoritmos de seleção de atributos para agrupamento de dados. O design fatorial experimental utilizado para gerar os conjuntos de dados produziu 324 conjuntos de dados considerando as replicações para melhorar a confidência estatística. Dois algoritmos de agrupamento de dados (k-means e EM) e quatro algoritmos de seleção de atributos foram sistematicamente aplicados para conjunto de dados. O teste de Wilcoxon/Mann-Whitney foi utilizado para testar se as diferenças entre as médias dos critérios externos obtidos pelos algoritmos são diferentes tem significância estatística. Três analises foram realizadas: a primeira que verifica se os algoritmos de seleção de atributos apresentam qualidades diferentes de agrupamentos, avaliados pelos dois critérios de validação externa; a segunda que verifica se os algoritmos de seleção de atributos, que não necessitam do número de grupos a priori, conseguem identificar o número corretos de agrupamentos; finalmente, a terceira que verifica se os algoritmos de seleção de atributos conseguem eliminar os atributos ruidosos, selecionando somente os relevantes. Levando as três analises em consideração, os algoritmos de seleção de atributos que apresentaram os melhores resultados foram o *clustvarsel* e o *vscc*.

Em relação a trabalhos futuros, os conjuntos de dados devem ser aumentados para dados com grandes dimensões, tais como os banco de dados de genes. Também devem ser aplicados conjuntos de dados reais, para verificar a validade de tais algoritmos. Assim, os conjuntos de dados sinteticos em conjunto com base de dados reais devem ser utilizados para compor um benchmark para novos algoritmos de seleção de atributos. Além disso, um número maior de algoritmos de seleção de atributos devem ser considerados.

[3]O algoritmo *sparseKmeans* não estima o número de agrupamentos, e o *SS* não foi empregado por não ser computacionalmente eficiente

Referências

[1] Hussein Almuallim e Thomas G Dietterich. "Learning with Many Irrelevant Features." Em: *AAAI*. Vol. 91. Citeseer. 1991, pp. 547–552.

[2] Jeffrey L Andrews e Paul D McNicholas. "Variable Selection for Clustering and Classification". Em: *Journal of Classification* (2013), pp. 1–18.

[3] Manoranjan Dash e Huan Liu. "Feature selection for classification". Em: *Intelligent data analysis* 1.3 (1997), pp. 131–156.

[4] Nema Dean e Adrian E Raftery. "The clustvarsel package". Em: *R package version 0.2-4* (2006).

[5] Arthur P Dempster, Nan M Laird e Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". Em: *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), pp. 1–38.

[6] Jennifer G Dy e Carla E Brodley. "Feature selection for unsupervised learning". Em: *The Journal of Machine Learning Research* 5 (2004), pp. 845–889.

[7] Brian S Everitt, Sabin Landau e Morven Leese. "Clustering analysis". Em: *Arnold, London* (2001).

[8] Chris Fraley e Adrian E Raftery. "Model-based clustering, discriminant analysis, and density estimation". Em: *Journal of the American Statistical Association* 97.458 (2002), pp. 611–631.

[9] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 1990.

		A	B	C	D	E	F
clustvarsel	A	0.000	0.011	0.013	0.051	0.075	0.319
EM	B	-0.011	0.000	0.002	0.039	0.063	0.307
vscc	C	-0.013	-0.002	0.000	0.037	0.061	0.305
sparseKmeans	D	-0.051	-0.039	-0.037	0.000	0.024	0.268
k-means	E	-0.075	-0.063	-0.061	-0.024	0.000	0.244
SS	F	-0.319	-0.307	-0.305	-0.268	-0.244	0.000
	mean	0.939	0.927	0.925	0.888	0.864	0.619

Table 3: Valores Médios e suas diferenças para o critério externo ARI

[10] Guojun Gan, Chaoqun Ma e Jianhong Wu. *Data clustering: theory, algorithms, and applications*. Vol. 20. Siam, 2007.

[11] Isabelle Guyon e André Elisseeff. "An introduction to variable and feature selection". Em: *The Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[12] Maria Halkidi, Yannis Batistakis e Michalis Vazirgiannis. "On clustering validation techniques". Em: *Journal of Intelligent Information Systems* 17.2-3 (2001), pp. 107–145.

[13] John A Hartigan e Manchek A Wong. "Algorithm AS 136: A k-means clustering algorithm". Em: *Applied statistics* (1979), pp. 100–108.

[14] Trevor Hastie et al. *The elements of statistical learning*. Vol. 2. 1. Springer, 2009.

[15] Eduardo R Hruschka e Thiago F Covoes. "Feature selection for cluster analysis: an approach based on the simplified Silhouette criterion". Em: *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*. Vol. 1. IEEE. 2005, pp. 32–38.

[16] Lawrence Hubert e Phipps Arabie. "Comparing partitions". Em: *Journal of classification* 2.1 (1985), pp. 193–218.

[17] Anil K Jain e Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[18] Robert E Kass e Adrian E Raftery. "Bayes factors". Em: *Journal of the american statistical association* 90.430 (1995), pp. 773–795.

[19] Leonard Kaufman e Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[20] Ron Kohavi e George H John. "Wrappers for feature subset selection". Em: *Artificial intelligence* 97.1 (1997), pp. 273–324.

[21] Huan Liu e Lei Yu. "Toward integrating feature selection algorithms for classification and clustering". Em: *Knowledge and Data Engineering, IEEE Transactions on* 17.4 (2005), pp. 491–502.

[22] James MacQueen et al. "Some methods for classification and analysis of multivariate observations". Em: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. California, USA. 1967, pp. 281–297.

[23] Paul David Mcnicholas e Thomas Brendan Murphy. "Parsimonious Gaussian mixture models". Em: *Statistics and Computing* 18.3 (2008), pp. 285–296.

[24] Glenn W Milligan e Martha C Cooper. "An examination of procedures for determining the number of clusters in a data set". Em: *Psychometrika* 50.2 (1985), pp. 159–179.

[25] Weiliang Qiu e Harry Joe. "Generation of random clusters with specified degree of separation". Em: *Journal of Classification* 23.2 (2006), pp. 315–334.

[26] Adrian E Raftery e Nema Dean. "Variable selection for model-based clustering". Em: *Journal of the American Statistical Association* 101.473 (2006), pp. 168–178.

[27] Gideon Schwarz et al. "Estimating the dimension of a model". Em: *The annals of statistics* 6.2 (1978), pp. 461–464.

[28] Mario F Triola et al. *Elementary statistics*. Pearson/Addison-Wesley, 2006.

[29] Daniela M Witten e Robert Tibshirani. "A framework for feature selection in clustering". Em: *Journal of the American Statistical Association* 105.490 (2010).

APPENDIX

O apêndice traz as mesmas informações da Tabela 3, mas para o coeficiente de Jaccard, Tabela 4.

		A	B	C	D	E	F
clustvarsel	A	0.000	0.011	0.008	0.049	0.074	0.289
EM	B	-0.011	0.000	-0.002	0.038	0.063	0.278
vscc	C	-0.008	0.002	0.000	0.041	0.066	0.281
sparseKmeans	D	-0.049	-0.038	-0.041	0.000	0.025	0.239
k-means	E	-0.074	-0.063	-0.066	-0.025	0.000	0.214
SS	F	-0.289	-0.278	-0.281	-0.239	-0.214	0.000
	mean	0.932	0.921	0.924	0.882	0.857	0.642

Table 4: Valores Médios e suas diferenças para o critério externo Jaccard