Medidas de Qualidade Aplicadas aos Algoritmos de Agrupamento da Shell Orion Data Mining Engine

Maicon Bastos Palhano[1], Allan Januário Ramos[1], Gabriel Felippe[1], Ruano Marques Pereira[1], Pedro Arns Junior[1], Ana Claudia Garcia Barbosa[1], Merisandra Côrtes de Mattos Garcia[1]

[1]Grupo de Pesquisa em Inteligência Computacional Aplicada - Curso de Ciência da Computação – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma, SC – Brasil

{maicon.palhano, gabrielheavy, ruanopereira}@hotmail.com, allanrjramos@gmail.com ,{agb, mem}@unesc.net

Resumo. *A evolução da tecnologia da informação possibilitou que grandes quantidades de dados pudessem ser processadas e armazenadas em bases de dados, porém quanto maior a quantidade, maior a complexidade de análise. O data mining surge como uma opção para analisar esses grandes conjuntos de dados de forma automática, buscando padrões e relações nos dados. Sendo composto por tarefas e métodos, onde uma dessas tarefas é a de agrupamento que gera clusters de dados. Porém, os clusters gerados por estes algoritmos devem ser avaliados por medidas de qualidade. Este artigo demonstra a aplicação de sete algoritmos de agrupamento em três bases de dados, a fim de analisar as partições geradas por meio de cinco medidas de qualidade.*

Palavras-chave*: Data Mining, Agrupamento, Shell Orion, Medidas de Qualidade.*

Abstract. *The development of information technology has enabled large quantities of data can be processed and stored in a database, but the greater the amount, the greater the complexity of analysis. Data mining emerges as an option for analyzing these large data sets automatically, looking for patterns and relationships in data. Comprised of tasks and methods, where one of these tasks is clustering which generates clusters of data. However, the clusters generated by these algorithms should be evaluated by measures of quality. This article demonstrates the application of seven clustering algorithms in three databases in order to analyze the partitions generated by five quality measures.*

Keywords*: Data Mining, Clustering, Shell Orion, Measures of Quality.*

1. Introdução

O expansivo crescimento dos bancos de dados empresariais, governamentais e científicos acarretou a incapacidade humana de interpretar e analisar essas informações, de modo que fossem necessários novos métodos e ferramentas capazes de analisar e extrair informações úteis a partir desses dados armazenados [Fayyad, Piatetsky-Shapiro e Smith 1996]. Sendo assim, de acordo com Sassi (2006) o *data mining* explora

repositórios de dados à procura de padrões e relacionamentos implícitos, encontrando dados que possam prever tendências ou comportamentos futuros.

O *data mining* tem sido aplicado em diversas áreas, com a finalidade, tanto de descrever características do passado, quanto predizer tendências futuras, dentre essas, destacam-se áreas como: finanças, energia, saúde, recursos hídricos, genética e biologia [Han e Kamber 2006].

A etapa de data mining é composta por tarefas e métodos, ambos possuindo suas particularidades, onde a escolha adequada de cada um irá depender do conhecimento acerca do domínio de aplicação do conjunto de dados utilizado [Kantardzic 2011].

Dentre as tarefas, tem-se a classificação, associação, sumarização, regressão, previsão de séries temporais e agrupamento [Goldschmidt e Passos 2005].

O agrupamento, segundo Han e Kamber (2006), reúne um conjunto de dados em grupos de objetos similares, formando *clusters*, buscando maximizar a semelhança entre objetos do mesmo *cluster* e minimizar a semelhança entre *clusters* [Larose 2005].

Agrupamento é uma tarefa onde não há grupos pré-definidos e exemplos que mostrem se os *clusters* encontrados pelos algoritmos são totalmente válidos [Halkidi et al 2002].

Mediante isso, a fim de se identificar a qualidade dos resultados gerados por algoritmos de agrupamento quanto a coesão e separação dos *clusters*, deve-se empregar medidas de qualidade voltadas a essa tarefa de data mining [Gan, Ma e Wu 2007].

Esta pesquisa enfatiza uma revisão de literatura na área de medidas de qualidade, destacando-se aquelas aplicadas no agrupamento, bem como a análise dos resultados gerados pelas medidas de qualidade Coeficiente de Partição, Coeficiente de Partição Entrópica, Xie-Beni, Dunn e C-Index nos algoritmos *Fuzzy C-Means* (FCM), *Robust C-Prototypes* (RCP), *Unsupervised Robust C-Prototypes* (URCP), *Density-based Spatial Clustering of Applications With Noise* (DBSCAN), *Ant Based Clustering* (ABC), *Adaptative Ant-Clustering Algorithm* (A²CA) e *Standard Ant Clustering Algorithm* (SACA) da tarefa de agrupamento da ferramenta *Shell Orion Data Mining Engine*, avaliando os resultados das suas partições, empregando-se três bases de dados.

2. Agrupamento

De forma geral, a maioria dos autores [Rezende 2005, Mary e Kumar 2012, Larose 2005] trazem definições muito parecidas com relação ao agrupamento, onde geralmente, sustentam como sendo a busca por *clusters* em uma base de dados de forma automática, por meio da similaridade entre os dados, sendo uma tarefa não-supervisionada e descritiva.

Rezaee et al (1998) salientam que, como o agrupamento é um processo não-supervisionado, onde não se tem conhecimento *a priori,* é inevitável o uso de algum tipo de avaliação das partições geradas, como medidas de qualidade.

3. Medidas de Qualidade no Agrupamento

O particionamento de um conjunto de dados resultante da aplicação de um algoritmo de agrupamento pode ser tradicional, com valores {0,1} ou *fuzzy*, com graus de pertinência variando entre [0,1] [Silva e Gomide 2004].

A divisão de um conjunto de dados gerada por algoritmos de agrupamento, geralmente não leva em consideração se essa estrutura realmente existe, tornando assim, importante a busca em saber se esses padrões encontrados são válidos, ou se apenas um particionamento sem significado foi gerado pelo algoritmo.

Após a execução dos algoritmos de agrupamento, é o momento em que podem surgir algumas dúvidas aos usuários relacionadas ao particionamento, como por exemplo: a) quantos *clusters* possuem o conjunto de dados; b) os resultados representam os dados reais; c) a maneira usada foi a melhor para particionar o conjunto de dados [Halkidi, Batistakis e Vazirgiannis 2001].

Considerando isso, tem-se a necessidade de métodos capazes de responder a estas perguntas, qualificando o particionamento gerado pelos algoritmos de agrupamento, de modo que se possa, além de avaliar o particionamento dos dados, identificar a quantidade exata de *clusters* presentes em um conjunto de dados, dados estes, que geralmente são não rotulados.

Geralmente, essas medidas no agrupamento, são definidas pela combinação da coesão e separabilidade dos *clusters*, no qual, coesão refere-se as medidas de proximidade dos conjuntos, como por exemplo, a variância; e separabilidade indica o quão distintos são os dois *clusters*, como a distância entre objetos de dois *clusters*, por exemplo [Rendón et al 2011].

Após o entendimento das medidas de qualidade no agrupamento, realizou-se uma pesquisa na literatura das principais medidas existentes, do período de 1973 a 2013: Dunn (1973), Variance Ratio Criterion (VRC) (1974), Coeficiente de Partição Entrópica (PE) Coeficiente de Partição (CP) (1974), Gamma (1975), C-Index (1976), Performance da Nebulosidade (1978), Davies-Bouldin (DB) (1979), Silhouette (1987), Fukuyama-Sugeno (FS) (1989), Xie-Beni (1991), Variações do Davies-Bouldin e Dunn, baseados em teoria dos grafos (1997), Separação e Dispersão (SD) (2000), S_Dbw (2001), Separação da Partição (2001), Contraste entre classes (2002), CS (2004), PBM (2004), Davies-Bouldin* (2005), Score Function (2007), Sym (2008), Distância Baseada em Pontos Simétricos (2009), COP (2010), Negentropy increment (2010), SV (2010) e OS (2010).

Uma das dificuldades encontradas em trabalhos relacionados à validação de *clusters* é a necessidade de um *framework* para interpretação de uma medida. Considerando-se que o valor final encontrado por uma medida seja, por exemplo, o número 10, qual o significado desse valor, ele é bom, ruim ou aceitável [Tan, Steinbach e Kumar 2005].

4. Análise das Medidas de Qualidade Aplicadas aos Algoritmos de Agrupamento da Shell Orion Data Mining Engine

A pesquisa desenvolvida consiste na análise de medidas de qualidade para validação dos resultados de algoritmos de agrupamento da *Shell Orion Data Mining Engine*[1], tendo em vista a importância destas medidas para a avaliação dos resultados gerados por esses algoritmos.

Após verificar que grande parte dos trabalhos publicados com relação a medidas de qualidade são internacionais, ressalta-se o quão pouco a área de medidas de qualidade para agrupamento é explorada em pesquisas nacionais, destacando a contribuição significativa do presente trabalho nesta área.

4.1. Bases de dados utilizadas

Para a aplicação dos algoritmos foram escolhidas as bases de dados *Iris*, *Wine* e das bacias hidrográficas.

As duas primeiras são conjuntos de dados obtidos a partir do UCI *Machine Learning Repository*[2], no qual consiste em um repositório de dados reais administrado pela Universidade da Califórnia. A terceira base de dados utilizada no trabalho refere-se a dados locais da região de Criciúma. Essa base de dados possui registros das bacias hidrográficas afetadas por poluentes provenientes da extração do carvão mineral na região, a base foi fornecida pelo professor Dr. Carlyle Torres de Bezerra Menezes do curso de Engenharia Ambiental da UNESC.

4.1.1. Base de dados da *Iris*

A base de dados da Iris refere-se às flores da família Iridáceas, com predominância nas espécies Iris-Setosa, Iris-Versicolor e Iris-Virginica. O conjunto de dados é composto por 150 registros, distribuídos em 50 registros para cada espécie e constituída por quatro atributos.

4.1.2. Base de dados da *Wine*

A base de dados *Wine* corresponde do resultado de análises químicas para determinar a origem de vinhos cultivados em uma região da Itália, porém provenientes de diferentes cultivadores de videiras, sendo composta por 178 registros e 13 atributos.

4.1.3. Base de dados das Bacias Hidrográficas

Esta base de dados corresponde a índices que trazem dados com relação a qualidade da água de três bacias hidrográficas da região de Criciúma. Os dados com indicadores ambientais somam 1723 registros e 20 atributos.

---

[1] Ferramenta acadêmica de *data mining* desenvolvida pelo Grupo de Pesquisa de Inteligência Computacional Aplicada desde 2005, contém até o presente momento as tarefas de associação, classificação e agrupamento.

[2] Repositório de dados disponível em: http://archive.ics.uci.edu/ml/

4.2. Seleção das medidas de qualidade e algoritmos utilizados

Dentre os algoritmos de agrupamento da Shell Orion, foram selecionados sete algoritmos para a aplicação das três bases de dados citadas anteriormente, sendo estes: FCM, RCP, URCP, DBSCAN, SACA, A²CA e ABC.

Com relação às medidas de qualidade analisadas, foram consideradas cinco medidas: Dunn, C-Index, Xie-Beni, Coeficiente de Partição e Coeficiente de Partição Entrópica, além da ampla aplicabilidade destas medidas em trabalhos da área, outro fator relevante foi o fato das mesmas já estarem implementadas nos algoritmos da *Shell Orion*, sendo os algoritmos FCM, RCP e URCP implementadas as medidas CP, CE e Xie-Beni e nos algoritmos DBSCAN, SACA, A²CA, ABC as medidas Dunn e C-Index.

A fim de uma melhor compreensão dos valores obtidos pelas medidas empregadas neste trabalho, as mesmas são descritas abaixo:

a) **Coeficiente de partição:** no intervalo de [1/C, 1], quanto mais próximo de um o valor da medida, mais compactos e bem separados serão os *clusters*;
b) **Coeficiente de partição entrópica:** esta medida é o contrário do coeficiente de partição, onde, quanto mais próximo de zero melhor será a qualidade dos *clusters*;
c) **Dunn:** esta medida encontra *clusters* compactos e separados quando valores maximizados são encontrados pela mesma;
d) **C-index:** diferentemente das medidas anteriores a C-Index avalia a coesão de cada *cluster* e não a separação de todos os *cluster*, onde no intervalo de [0,1], quanto mais próximo de zero, mais bem compacto é o *cluster* avaliado.

5 Resultados Obtidos

Todos os algoritmos foram executados com as três bases de dados, porém, alguns não obtiveram resultados satisfatórios quanto a separação ou a coesão dos *clusters* gerados, um dos fatores, foi o fato de algumas bases de dados possuírem *clusters* muito próximos afetando ou o desempenho dos algoritmos ou das medidas consideradas.

5.1. Algoritmos versus base de dados *Iris*

A base de dados da *Iris* possui três *clusters*, porém dois *clusters* são muito próximos e quase sobrepostos, devido a isso alguns autores da literatura aceitam a presença de dois *clusters* ao invés de três. Pela Tabela 1 é possível observar que CP e CE indicam a presença de três *clusters* nos algoritmos RCP e URCP, no algoritmo FCM obteve melhores resultados em dois *clusters*, com relação ao XB, o mesmo indicou a presença de dois *clusters* pelo algoritmo FCM e RCP, no algoritmo URCP encontrou um valor ruim.

Tabela 1 – *Iris*: Medidas de qualidade algoritmos FCM, RCP e URCP

Alg.	2 *Clusters*			3 *Clusters*			4 *Clusters*		
	CP	**CE**	**XB**	**CP**	**CE**	**XB**	**CP**	**CE**	**XB**
FCM	0,6824	0,4936	0,2120	0,54481	0,7880	0,34001	0,4302	1,0649	0,3153
RCP	0.6518	0.5213	0.1051	0.9496	0.1217	3.4356	0.4287	1.054.7	0.2741
URCP	X	X	X	0.7813	0.5487	9,8947	X	X	X

Pela Tabela 2 pode-se visualizar que o Dunn indicou um valor ruim para todos os algoritmos, porém a C-Index indicou valores contrários apontando *clusters* compactos, tendo em vista que o C-Index avalia cada *cluster* de forma individual.

Tabela 2 – *Iris*: Medidas de qualidade algoritmos DBSCAN, ABC, SACA, A²CA

Algoritmos	Nº do *Cluster*	Dunn	C-Index
DBSCAN	1	0.344837	0.047277
	2	0.344837	0.103562
	3	0.344837	0.130308
ABC	1	0.012217	0.198707
	2	0.012217	0.270436
	3	0.012217	0.227525
SACA	1	0.023212	0.069286
	2	0.023212	0.151758
A²CA	1	0.053226	0.032541
	2	0.053226	0.164694

5.2. Algoritmos versus Base de dados *Wine*

A base de dados *Wine*, segundo a literatura possui três *clusters*, porém devido a sua alta dimensionalidade alguns algoritmos não a conseguem particioná-la corretamente. Na Tabela 3 pode-se observar que todas as medidas obtiveram os melhores valores em dois *clusters*, indicando como tal o número ideal de *clusters*.

Tabela 3. *Wine:* Medidas de Qualidade algoritmos FCM, RCP e URCP

Alg.	2 *Clusters*			3 *Clusters*			4 *Clusters*		
	CP	CE	XB	CP	CE	XB	CP	CE	XB
FCM	0,7152	0,4456	0,1974	0.5723	0.7456	0.2477	0.5070	0.9344	0.2136
RCP	0.5039	0.6891	0.1432	0.3850	0.1217	1.0180	0.2668	1.3488	0.03495
URCP	0.7883	0.3415	1.4340	X	X	X	X	X	X

Pela Tabela 4 observa-se que o Dunn mostrou uma bom agrupamento, pois foi encontrado valores bons para todos os algoritmos, exceto o DBSCAN que não conseguiu agrupar a base da *Wine* e o C-Index indicou *clusters* compactos para todos os algoritmos exceto o DBSCAN, assim como o Dunn.

Tabela 4. *Wine*: Medidas de Qualidade algoritmos DBSCAN, ABC, SACA e A²CA

Algoritmos	Nº do *Cluster*	Dunn	C-Index
DBSCAN	X	X	X
ABC	1	3.612698	0.277684
	2	3.612698	0.426757
	3	3.612698	0.422466
SACA	1	3.469666	0.345134

	2	3.469666	0.247756
A²CA	1	3.955415	0.416503
	2	3.955415	0.271009
	3	3.955415	0.289912

5.3. Algoritmos versus base de dados das Bacias Hidrográficas (BH)

Os dados desta base de dados foram aplicados em alguns trabalhos do Grupo de Inteligência Computacional Aplicada[3], porém em nenhum deles chegou a conclusão do número ideal de *clusters*, cada trabalho resulta em um agrupamento diferente. Na Tabela 5 pode-se visualizar que as medidas CP, CE encontraram os melhores valores em todos os algoritmos em dois *clusters*, XB indicou a presença de quatro *clusters* pelo algoritmo FCM e valores ruins para os algoritmos RCP e URCP.

Tabela 5. BH: Medidas de Qualidade algoritmos FCM, RCP e URCP

Alg.	**2 *Clusters***			**3 *Clusters***			**4 *Clusters***		
	CP	**CE**	**XB**	**CP**	**CE**	**XB**	**CP**	**CE**	**XB**
FCM	0.7658	0.3831	0.1627	0.6826	0.5619	0.2059	0.64424	0.6821	0.0949
RCP	0.6997	0.4612	12.1975	0.6231	0.6470	22.1062	X	X	X
URCP	0.7425	0.4006	35.9647	X	X	X	X	X	X

Pela tabela 6 observa-se que os algoritmos DBSCAN e o ABC não conseguiram agrupar esta base de dados, com relação ao Dunn nos algoritmos SACA e A²CA obteve bons valores bem como o C-Index.

Tabela 6. BH: Medidas de Qualidade algoritmos DBSCAN, ABC, SACA e A²CA

Algoritmos	**Nº do *Cluster***	**Dunn**	**C-Index**
DBSCAN	X	X	X
ABC	X	X	X
SACA	18	1.790367	0.0250 - 0.0962
A²CA	13	7.0066	0.0187 - 0.0991

Observa-se na aplicação dos sete algoritmos, que nem todos conseguiram atender os requisitos mínimos de agrupamento, pois quando algumas bases de dados possuíam: muitas dimensões (*Iris* e *Wine*), algum tipo de ruído ou com muitos elementos ou atributos como a das bacias hidrográficas, alguns algoritmos, como foi o caso do DBSCAN, ABC e SACA não conseguiram particionar corretamente os *clusters*.

Com relação às cinco medidas de qualidade aplicadas pode-se afirmar que as medidas que melhor mostraram o número ideal de *clusters* e se estes eram compactos e bem separados, sem contrariedades foram o Coeficiente de Partição, Coeficiente de Partição Entrópica e C-Index. Já as medidas que mais foram arbitrárias em seus resultados foram a Xie-Beni e Dunn.

[3] Grupo de pesquisa do curso de Ciência da Computação.

6. Considerações Finais

Este artigo mostrou uma pesquisa realizada das medidas de qualidade existentes e empregaram-se sete algoritmos de agrupamento da Shell Orion, sendo cada um executado com três bases de dados diferentes. O particionamento realizado pelos algoritmos para cada base de dados foi analisado por meio de medidas de qualidade, chegando-se a conclusão de que a maioria dos algoritmos da Shell Orion conseguiram identificar corretamente os *clusters* quando a base de dados era pequena, porém com bases de dados maiores como foi o caso das bacias hidrográficas e *Wine*, alguns algoritmos não conseguiram agrupar de forma correta os dados.

Com relação às medidas de qualidade observou-se que estas são muito relativas, pois nem sempre uma medida que funciona corretamente em uma base de dados ou algoritmo específico se comportará deste mesmo modo com outra base de dados ou algoritmo. Destaca-se também a vulnerabilidade das medidas com relação a ruídos, bases de dados com *clusters* sobrepostos, e bases de dados com múltiplas dimensões.

Com isso, conclui-se que medidas de qualidade no agrupamento é uma área relativamente nova e ainda está em constantes estudos e buscando cada vez mais novas medidas que se comportem melhor a uma determinada aplicação, pois não existe uma medida universal que consiga atender a todo algoritmo ou base de dados, porém é a única forma de auxiliar o usuário quanto ao número ideal de *clusters* e a medir a qualidade dos *clusters* gerados.

Referências

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic.From data mining to knowledge discovery in databases. AI Magazine, Providence, v.17, n. 3, pp. 37-54, autumn 1996.

GAN, G., MA, C. e WU, J. **Data clustering**: theory, algorithms and applications. Philadelphia: SIAM, 2007.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data mining**: uma guia prático: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment using multi representatives. Poster paper in the Proceedings of SETN Conference, April 2002, Thessaloniki, Greece

HALKIDI, M.; VAZIRGIANNIS, M. Clustering Validity Assessment: Finding the optimal partitioning of a data set. In the Proceedings of ICDM Conference, California, USA, November 2001 [Halkidi et al 2002].

HAN, J., KAMBER, M., **Data Mining Concepts and Techniques**. Morgan Kaufman Publishers, San Francisco, USA, 2006.

KANTARDZIC, M. **Data mining:** Concepts, models, methods, and algorithms. New York, NY: Segunda Edição, John Wiley and Sons, 2011.

LAROSE, Daniel T. **Discovering knowledge in data: an introduction to data mining.** Hoboken: Wiley-Interscience, 2005.

MARY, Angel Latha S.; KUMAR, K.R. Shankar. Density Based Dynamic Data Clustering Algorithm based on Incremental Dataset- *Journal of Computer Science* 8 (5): pp. 656-664, 2012.

RENDÓN, Eréndira et al. Internal versus External cluster validation indexes. International Journal Of Computers And Communications, Issue 1, Vol. 5, pp. 27-33, 2011.

REZAEE, Babak, A cluster validity index for fuzzy clustering, *Fuzzy Sets and Systems*, Volume 161, Issue 23, 1 December 2010, pp. 3014-3025.

REZENDE, Solange Oliveira. **Sistemas inteligentes:** fundamentos e aplicações. Barueri: Manole, 2005.

SASSI, Renato José. **Uma arquitetura híbrida para descoberta de conhecimento em bases de dados:** teoria dos rough sets e redes neurais artificiais mapas auto organizáveis.2006. 169 f. Tese (Doutorado em Sistemas Eletrônicos) – Escola Politécnica, Universidade de São Paulo. São Paulo, 2006.

SILVA, Leila R. S. da; GOMIDE, Fernando. Um estudo comparativo entre as funções de validação para agrupamento nebuloso de dados- IV Congresso Brasileiro de Computação, Itajaí, 2004.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin; **Introdução ao Data Mining Mineração de dados**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.