ANÁLISE DE PARÂMETRO PARA OTIMIZAÇÃO DE MEDIDAS RELACIONADAS A AGRUPAMENTOS DE DADOS

B. C. Ernani, M. T. Selma, P. A. Adriano
Biolab, Faculdade de Engenharia Elétrica, Universidade Federal de Uberlândia, Uberlândia – Minas Gerais, Brasil,
e-mail: ernani@iftm.edu.br

***Resumo* - A técnica de análise de agrupamento (*clustering analysis*) é uma ferramenta importante na pesquisa científica, podendo ser utilizada em diversas áreas do conhecimento tais como medicina, biologia e estatística. Agrupar dados é uma forma de refletir a estrutura interna dos dados e identificar classes presentes nesses agrupamentos, de modo que haja homogeneidade dentro das mesmas classes e heterogeneidade entre classes diferentes. Existem três métodos de agrupamentos utilizados para encontrar o particionamento ótimo que são: os métodos hierárquicos, baseados em teorias dos grafos e baseados em função objetivo. Neste trabalho foi utilizado o Algoritmo baseado em função objetivo *Fuzzy C-Means* em conjunto com a técnica de reamostragem *bootstrap*. A idéia é variar o índice de nebulosidade para encontrar a melhor faixa de valores a ser utilizada para a classificação dos dados e consequentemente obtenção de melhores particionamentos. A qualidade de comparação é baseada em medidas de comparação tradicionais tais como Hubert, Jaccard, F1. As bases de dados utilizadas foram a *Iris*, *Wine* três bases de dados artificiais. Os resultados obtidos até o momento demonstram que a melhor faixa de valor para o índice de nebulosidade está entre 1.07 e 1.2 para os índices de medidas adotas.**

***Palavras-Chave* – Agrupamento de Dados, Índice de Nebulosidade, *Fuzzy C-Means*.**

ANALYSIS FOR PARAMETER OPTIMIZATION OF MEASURES RELATING OF DATA CLUSTERING

***Abstract* - The technique of clustering analysis is an important tool in scientific research, can be used in various fields of knowledge such as medicine, biology and statistics. Group data in clusters is a way to reflect the internal data structure and identify classes present in this clusters, so that there is within the same homogeneity and heterogeneity between different classes. There are three types of clustering methods used to find optimal partitioning: hierarchical methods, based on graph theory and based on objective function. In this study we used the objective function algorithnm based on Fuzzy C-Means and also the bootstrap resampling technique. The ideia is to vary the index cloudiness in order to find the best value to be used for sorting the data and thus obtain better partitioning. The comparison quality is based on traditional measures of comparition such as Hubert, F1 and F1. The databases used were the Iris, Wine and three artificial databases. The results obtained so far show that the best range for the index of cloudiness is between 1.07 and 1.2 for teh contents of measures adotas.**

1

***Keywords* - Clustering Analysis, weighting exponent, index cloudiness, *Fuzzy C-Means* .**

I. INTRODUÇÃO

A humanidade vive hoje em um mundo repleto de dados. Todos os dias, as pessoas encontram uma grande quantidade de informações e as armazenam para posterior análise e gestão.

Uma das ferramentas que podem ser utilizadas para estudo desses dados é a análise de agrupamentos (*clustering analysis*), que é uma das mais antigas técnicas em que não são feitas suposições com relação ao número de grupos ou à estrutura existente dentro do grupo. Os procedimentos exploratórios são frequentemente úteis para o entendimento da natureza complexa, existente nas relações multivariadas. Buscar nos dados uma estrutura de agrupamentos naturais é uma importante técnica exploratória, pois agrupamentos podem fornecer um meio informal para acesso à dimensionalidade, identificando tendências e sugerindo hipóteses relativas às semelhanças [1].

A partir das técnicas de agrupamento, diversos estudos a aplicam como uma ferramenta para análise, tendo sido utilizada nas áreas de processamento de imagens, biologia, reconhecimento de dados, mineração de dados, sensoriamento remoto, bioinformática, dentre outras [2].

Existem três métodos de agrupamentos que são utilizados para encontrar o particionamento ótimo: métodos "hierárquicos", métodos baseados em "teoria dos grafos" e métodos baseados em "função objetivo". Os métodos baseados em função objetivo são muito utilizados [7] e dentre eles pode-se destacar a Classificação Nebulosa (*Fuzzy*).

Independentemente do algoritmos de classificação propostos [3], os resultados de todos devem ser validados de forma quantitativa e objetiva. Uma das maneiras existentes para a validação é tomar como base as estabilidade da solução encontrada[4]. Esse critério de validação utiliza

exclusivamente a própria base de dados, isso implica dizer que deve haver a reamostragem dos dados.

Dessa forma, a reamostragem pode ser utilizada para escolha do particionamento mais consistente presente na base de dados. A idéia é realizar a comparação entre a base de dados completa e sub-amostras dessa base de dados. Espera-se que, quando o número de grupos estiver correto, as sub-amostras e a base de dados original tenham a mesma estrutura de grupos. Para um número incorreto de grupos o resultado do agrupamento deve ser instável (minimizando os valores das medidas utilizadas) [4][5].

Porém, a estabilidade em uma Classificação *Fuzzy* não pode ser definida simplesmente pelo método de reamostragem, pois, nessa classificação existe um parâmetro chamado "expoente de ponderação ***m***" (também chamado de índice de nebulosidade ou índice de fuzificação), que é empírico e seu valor influencia substancialmente os resultados da classificação nebulosa[8].

Neste contexto, este trabalho realizou um estudo para buscar a melhor faixa de valores de ***m,*** ou seja, que leve aos agrupamentos mais estáveis. A qualidade de comparação é baseada em medidas de comparação tradicionais tais como Hubert, Jaccard e F1. Foram utilizadas as bases de dados *Iris*, *Wine* e três bases de dados artificiais.

II. AGRUPAMENTO DE DADOS

As pessoas sempre tentam buscar recursos que possam auxiliar na tomada de decisão. Assim, a análise de agrupamento de dados tende a agrupar dados com base na similaridade ou dissimilaridade (distância) de acordo com determinadas normas ou regras [9].

A análise de agrupamento (Figura 1) consiste basicamente em representação padrão (extração de características e/ou seleção), projeto de agrupamento (algoritmo de seleção, definição da uma medida de proximidade padrão apropriada para o domínio de dados), validação do agrupamento, interpretação dos resultados.

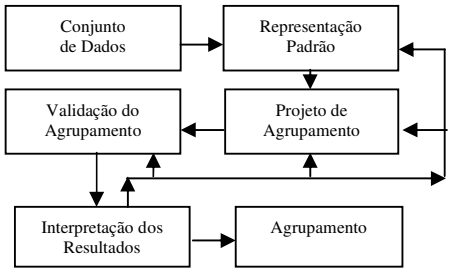

Fig. 1. Fases do agrupamento de dados (modificada de [9])

A "Representação padrão" extrai (seleciona) características distintas de um conjunto de objetos. São utilizadas algumas regras para gerar os novos dados a partir dos dados originais. Geralmente, as características devem conter padrões distintos pertencentes a diferentes grupos [9].

O "Projeto de agrupamento" é a combinação da seleção de uma medida de proximidade, a qual afeta diretamente a formação dos grupos e o desenvolvimento das rotinas de critério de semelhanças (agrupamentos), pois, quase todos os algoritmos de agrupamentos são explícita ou implicitamente ligados a alguma definição de medida de proximidade e não há um algoritmo universal para todos os problemas [9].

A "Validação do agrupamento" é responsável por avaliar a saída do procedimento de agrupamento, pois, a identificação de parâmetros ou a ordem de apresentação dos padrões de entrada podem afetar os resultados finais. Portanto, normas de avaliação e critérios são importantes para fornecer aos usuários resultados com alto grau de confiabilidade [9].

Finalizando a seqüência das fases do agrupamento, a "Interpretação dos resultados" deverá ser capaz de fornecer aos usuários, conhecimentos significativos a partir dos dados originais, possibilitando a tomada de decisões e solução de problemas [9].

III. LÓGICA FUZZY

Também conhecida como "Lógica Nebulosa" foi introduzida em 1965 pelo matemático, Lofti Asker Zadeh, por meio da publicação de um trabalho sobre Conjuntos *Fuzzy*, baseado na lógica multinível, o qual mostra o tratamento dos aspectos imprecisos e incertos. A lógica tradicional (binária ou *crisp*) trata os valores 0 e 1 (falso ou verdadeiro) e na lógica *Fuzzy* é possível encontrar os valores "entre 0 e 1" podendo ser quase falso como também quase verdadeiro, permitindo assim, explorar a tolerância à imprecisão, à incerteza e à veracidade parcial para alcançar tratabilidade, robustez [10].

Assim, a teoria do conjunto nebuloso diz que, dado um determinado elemento que pertence a um domínio, é verificado o grau de pertinência do elemento em relação ao conjunto. O grau de pertinência é a referência para verificar o quanto é possível esse elemento poder pertencer ao conjunto. Um conjunto nebuloso *A* na base de dados *X* é caracterizado por uma função de pertinência $\mu_A(x)$ a qual associa cada ponto em *X* a um número real no intervalo [0,1], com o valor de $\mu_A(x)$ representando o *grau de pertinência* de *x* em *A*, ou seja, *A* é um conjunto de pares ordenados do elemento genérico *x* dado por:

$$A = \left\{ \left(x, \mu_A(x)\right) \mid x \in X \right\} \qquad (1)$$

se $\mu_A(x) = 1$ temos pertinência total ao conjunto nebuloso *A*, e $\mu_A(x) = 0$, não temos pertinência ao conjunto nebuloso *A*. Um valor próximo de zero indica "baixo" grau de pertinência e um valor próximo de 1, indica "alto" grau de pertinência.

A Figura 2 exemplifica agrupamentos nebulosos e não nebulosos sendo os retângulos H_1 e H_2 agrupamentos *crisp* (*hard*) e as elipses F_1 e F_2 a saída do algoritmo nebuloso.

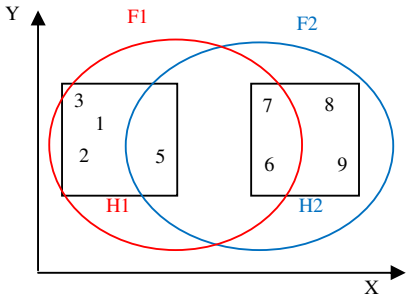

Fig. 2. Agrupamento nebuloso (modificada de [2]).

A. Algoritmo Fuzzy C-Means (FCM)

É um típico algoritmo de análise de grupos e tem sido amplamente utilizado na área científica [11]. Suas principais características são baixa complexidade computacional e facilidade na implementação [12][13]. É o equivalente nebuloso do algoritmo crisp *Fuzzy K-Means* [14].

O processo de cálculo do algoritmo FCM (Equação 2) é iterativo. Assim, o propósito é minimizar o índice de desempenho da pseudopartição nebulosa J_m (ou função objetivo) que mede a distância entre os centros dos grupos e os elementos dentro dos grupos, desta forma, quanto menor seu valor, mais otimizada estará a pseudo-partição ou partição nebulosa U.

$$J_m(U,V) = \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^m (d_{ij})^2 \qquad (2)$$

Onde:

U - Matriz de graus de pertinência nebulosos.

V - Conjunto de vetores que representa os c centros dos grupos.

c - Número de grupos, sendo inteiro, positivo e maior que 1 e menor que n (número de dados).

μ_{ij} - Pertinência do dado j no grupo i.

d_{ij} - Distância do dado j ao centro do grupo i.

m - Expoente de ponderação (índice de nebulosidade).

Basicamente os passos para execução do algoritmo *Fuzzy C-Means* para uma base de dados são: definir (aleatoriamente) o índice de nebulosidade m (maior ou igual a 1, tendendo ao infinito), estabelecer o número de agrupamentos c (maior que 1 e menor que a quantidade dados), definir o critério de parada, ou seja o número máximo de iterações t *(t = 0,1,2,3....)*, calcular vetor de protótipo (centros associados a cada partição de cada grupo ($V^{(t)}$)), atualizar os graus de pertinências (μ_{ij}).

Os agrupamentos obtidos pelo algoritmo FCM devem ser validados, pois como os demais algoritmos de agrupamento ele produz um modelo de particionamentos para uma base de dados, quer existam ou não. Essa avaliação pode ser feita verificando se o modelo de agrupamento obtido é o que mais se adéqua ao conjunto de dados ou avaliando-se a qualidade do agrupamento.

A reamostragem com reposição (*bootstrapping* [15]) pode ser usada para estimar índices relativos. A comparação é feita entre o agrupamento resultante para uma dada amostra da base de dados e um agrupamento de referência. Assim, os métodos baseados em reamostragem são utilizados para verificar onde a base de dados possui uma estrutura de agrupamentos ou onde o resultado é somente um artefato do algoritmo e também para selecionar o modelo do número de grupos [6][16].

Medidas de comparação binária tais como F1, Acc (Classificação cruzada), Diff (Diferença média quadrática), Jaccard, Índice Randômico, Hubert, *Folkes and Mallows*, Randaj (Randômico Ajustado) podem ser utilizadas para validação [17].

IV. MATERAIS E MÉTODOS

Os testes foram desenvolvidos em um *notebook* com processador Intel® core™ i5-2410M CPU@2.30 GHz, 6 GB de Memória RAM, com sistema operacional Windows 7 e software MATLAB®.

Neste trabalho foram feitas análises de cinco bases de dados: Artificial 1, Artificial2, Artificial3, *Iris* e *Wine*. As Bases de dados *Iris* e *Wine* são bases reais e foram obtidas no repositório *UCI Machine Learning Repository* [18].

A Tabela III apresenta os dados utilizados e seus respectivos números de grupos. As Figuras 3, 4, 5, 6 e 7 apresentam os gráficos das cinco bases de dados utilizadas.

TABELA III
Bases de Dados

Conjunto de Dados	Número de dados	Número de Atributos	Número de grupos
Artificial 1	400	2	4
Artificial 2	350	2	7
Artificial 3	1000	3	5
Iris	150	4	3
Wine	178	13*	3

* neste trabalho só foram utilizados os atributos 7, 10 e 13.

O método teve início com a normalização da base de dados (dados originais) para média zero e desvio padrão unitário, para remover os efeitos de escala.

Na execução do algoritmo FCM foi utilizados um número de iterações t = 100; índice de nebulosidade m = 1.01, 1.02, 1.03, 1.04, 1.05, 1.06, 1.07, 1.1, 1.2, 1.5, 1.8, 2.0, 2.3, 2.5, 2.8, 3.0, 3.5, 4.0 e 4.5; critério de parada $\varepsilon = \leq 0.00001$ e o número de grupos c = 2, 3, ..., 8.

O número de grupos escolhido, variou de 2 a 8, para permitir uma melhor análise em torno do número de agrupamentos existente nas bases de dados analisadas.

O método de reamostragem com reposição (*bootstrapping*) foi aplicado para geração da sub-amostra de dados contendo 90% dos dados da referência (base de dados original). Foram utilizadas 100 sub-amostras.

O algoritmo FCM foi então aplicado a cada sub-amostra, obtendo-se um modelo de agrupamento. Os centros desse modelo foram então reaplicados na base de dados original, obtendo-se assim, um modelo de grupos para a mesma.

A seguir, para cada sub-amostra, foram calculadas as medidas de comparação binária escolhidas para esse trabalho (Hubert, Jaccard e F1). O valor final para cada medida foi obtido aplicando-se a média entre os resultados das 100 sub-amostras.

O valor máximo obtido para cada medida, entre todos os particionamentos realizados, indica o número de grupos definido pela medida como sendo o mais correto (mais estável) para a base de dados em análise.

V. RESULTADOS E DISCUSSÃO

As Tabelas IV, V, VI, VII e VIII mostram as variações dos índices de nebulosidade m em relação às medidas de comparação Hubert (Hub), F1 e Jaccard.

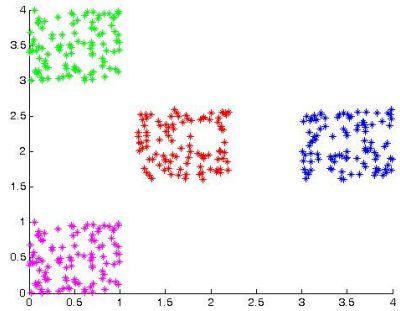

Fig. 3. Base de Dados Artificial 1 [17].

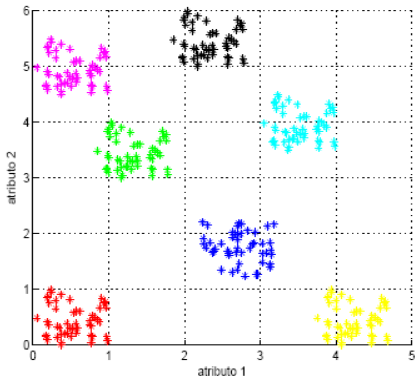

Fig. 4. Base de Dados Artificial 2 [17].

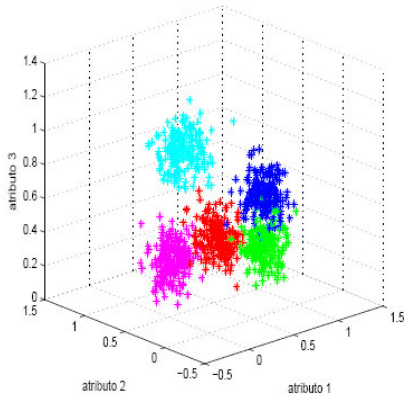

Fig. 5. Base de Dados Artificial 3 [17].

Nota-se que nas Tabelas IV e V, a variação do índice de nebulosidade, para faixa de valores baixos e/ou altos não apresentou variação significante na quantidade de seus grupos, mantendo o quantitativo de 4 e 7 grupos conforme base de dados original, por serem linearmente separados.

Já a Base de Dados Artificial 3, por não ter todos os grupos linearmente separados, observou-se uma pequena variação na quantidade de grupos durante a variação do índice de nebulosidade após a faixa 2.0, conforme tabela VI.

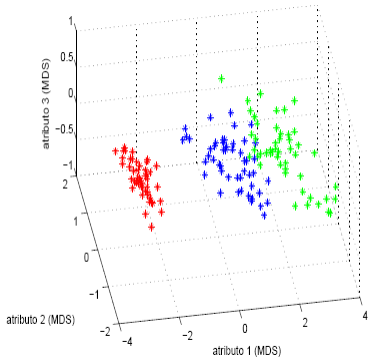

Fig. 6. Base de Dados *Iris* [17].

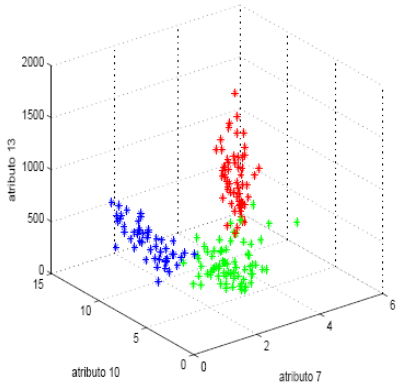

Fig. 7. Base de Dados *Wine* [17].

A Base de Dados *Iris*, nas medidas de comparação Hubert a partir do índice nebuloso 2.3 percebe-se um pequena variação na quantidade grupos para cima, isso devido a base ter um grupo totalmente separado e os outros dois parcialmente sobrepondo um ao outro.

A Base de Dados original *Wine* possui 3 grupos conforme Figura 7, os grupos não se encontram linearmente separados, existe uma proximidade entre eles, sendo alguns dados sobrepostos e outros dados mais distantes. Porém, na Tabela VIII pode-se observar que somente nas medidas de comparação Jaccard e F1, a partir do índice 2.0 houve variação, as demais mantém os três grupos.

TABELA IV
Variação do índice de nebulosidade
Base de Dados Artificial 1

m	Hub	cluster	Jaccard	cluster	F1	Cluster
1.01	0.9874	4	0.9812	4	0,0000	0
1.02	1.0000	4	1.0000	4	1,0000	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.0	0.7792	4	0.7101	4	0,8359	4
2.3	0.6781	4	0.5953	4	0,7242	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.5	0.4773	4	0.3770	4	0,5235	2
4.0	0.4501	4	0.3664	2	0,5439	2
4.5	0.4397	3	0.3627	2	0,5399	2

TABELA V
Variação do índice de nebulosidade
Base de Dados Artificial 2

m	Hub	cluster	Jaccard	cluster	F1	Cluster
1.01	0.9818	7	0.9710	7	0,0000	0
1.02	0.0962	7	0.9802	7	0,9586	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.0	0.7625	7	0.6581	7	0,7847	7
2.3	0.6530	7	0.5282	7	0,6813	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.5	0.4313	7	0.3686	2	0,5755	2
4.0	0.4130	4	0.3590	2	0,5567	2
4.5	0.4160	4	0.3526	2	0,5437	2

TABELA VI
Variação do índice de nebulosidade
Base de Dados Artificial 3

m	Hub	cluster	Jaccard	cluster	F1	cluster
1.01	0.9853	5	0.9780	5	0,0000	0
1.02	0.9845	5	0.9764	5	0,9929	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.0	0.5502	5	0.4910	3	0,7187	2
2.3	0.5044	5	0.4302	2	0,6631	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.5	0.4296	4	0.3663	2	0,5614	2
4.0	0.4246	3	0.3572	2	0,5448	2
4.5	0.4315	3	0.3515	2	0,5338	2

TABELA VII
Variação do índice de nebulosidade
Base de Dados *Iris*

M	Hub	cluster	Jaccard	cluster	F1	cluster
1.01	0.9859	2	0.9875	2	0,9953	2
1.02	0.9886	2	0.9899	2	0,9959	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.0	0.5213	2	0.6303	2	0,8237	2
2.3	0.4317	3	0.5504	2	0,7608	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.5	0.3926	3	0.4145	2	0,6167	2
4.0	0.3976	3	0.3932	2	0,5880	2
4.5	0.4075	3	0.3793	2	0,5683	2

TABELA VIII
Variação do índice de nebulosidade
Base de Dados *Wine*

M	Hub	cluster	Jaccard	cluster	F1	cluster
1.01	0,9094	3	0,8894	3	0,9768	3
1.02	0,9191	3	0,9013	3	0,9755	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.0	0,4564	3	0,4666	2	0,7169	2
2.3	0,3971	3	0,4248	2	0,6486	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.5	0,3862	3	0,3662	2	0,5566	2
4.0	0,3913	3	0,3565	2	0,5427	2
4.5	0,3994	3	0,3509	2	0,5332	2

VI. CONCLUSÕES

Este trabalho apresentou um estudo da variação do índice de nebulosidade, usando o algoritmo *Fuzzy C-Means*, que leve aos agrupamentos mais estáveis. Observou-se que para grupos bem definido e separados linearmente como o caso das Bases Artificiais 1, 2 e 3, predomina a faixa de valores do índice de nebulosidade entre 1.01 e 2.3 nas três medidas de comparação Hubert, F1 e Jaccard. Para a Base de Dados *Iris* nota-se que de 1.01 até 2.0 é a faixa ideal mantendo os grupos. Já a Base de Dados *Wine*, a variação do índice nebuloso apresentou variação somente na medida de comparação Jaccard e F1 a partir do valor de m = 2.0, não havendo alteração para as medidas de Hubert.

Observa-se que quando m varia entre 1.07 e 1.2 todos os índices de medida estimam corretamente o número de grupos, contrariando alguns trabalhos anteriores cuja faixa ideal do valor de m seria de 1,5 a 2,5. Essa variação em relação aos trabalhos anteriores, reafirma a importância da determinação correta do valor de m, pois a diferença encontrada pode ser devida aos índices de medida adotado, indicando que o valor de m é dependente desses índices.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] R. A. Johnson, *Applied Multivariate Statistical Analysis*, Prentice Hall, 3ª Edição, New Jersey, 1992.

[2] A. K. Jain, M. N. Murty, P. J. Flynn, *Data clustering: A review. ACM Computing Surveys (CSUR),* ACM Press, vol. 31, no. 3, pp. 264-323, New York, September 1999.

[3] A. K. Jain, R. C. Dubes, *Algorithms for Clustering Data*, Upper Saddle River, NJ, USA, Prentice Hall, 1988.

[4] V. Roth, T. Lange, M. L. Braun, J.M. Buhmann, *A resampling approach tocluster validation. In: 15th Computational Statistics (COMPSTAT)*, Physica-Verlag Heidelberg, Germany: Berlin 2002.

[5] C. Borgelt, *Resampling for fuzzy clustering. In: Proc. Symposium on Fuzzy Systems in Computer Science*, Otto-von-Guericke-Universitat, Magdeburg, Germany, 2006.

[6] C. Borgelt, *Prototype-based Classi_cation and Clustering*, Tese (Doutorado) - Otto-von-Guericke-Universitat, Magdeburg, Germany, November 2005.

[7] M. Sarkar, T. Y. Leong, *Fuzzy k-means clustering with missing values*, *In: American Medical Informatics Association Annual Symposium (AMIA)*, Medical Publishers, pp. 588-592, Philadelphia, 2001.

[8] N. R. Pal, J. C. Bezdek, "On cluster validity for the fuzzy c-means model". *IEEE Transactions on Fuzzy Systems, IEEE Computer Society, Washington, DC, USA*, vol. 3, no. 3, pp. 370-379, August 1995.

[9] R. Xu, D. C. Wunsch II, "Survey of clustering algorithms", *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678.

[10] L. A. Zadeh, *Fuzzy logic, neural networks and soft computing*. Comunnications ofthe ACM, ACM Press, vol. 37, no. 3, pp. 77-84, New York, USA, March 1994.

[11] D. Zhang, S. Chen, *Clustering incomplete data using kernel-based fuzzy c-means algorithm*, Neural Processing Letters, Kluwer Academic, vol. 18, no. 12, pp. 155-162, Netherlands, NY, USA, December 2003.

[12] A. G. D. Nuovo, V. Catania, S. D. Nuovo, S. Buono, "Envolving fuzzy c-means: An intelligent technique for efficient diagnosis of children mental retardation level from databases with missing values", *In: International Conference on Artificial Intelligence*, Las Vegas, USA, 2006.

[13] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*, John Wiley and Sons, Chichester, England, 1999.

[14] X. L. Xie, G. Beni, "A validity measure for fuzzy clustering". *IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Computer Society,* vol. 13, no. 8, pp. 841-847, Washington, November 1991.

[15] B. Efron, R. J. Tibshirani, *An Introduction to the Boosttrap*, Chapman and Hall, New York, 1993.

[16] M. H. Law, H. A. K. Jain, *Cluster Validity by Bootstrapping Partitions*. East Lansing, Michigan, 2003.

[17] S. T. Milagre, *Análise do Número de Grupos em Bases de Dados Incompletas Utilizando Agrupamentos Nebulosos e Reamostragem Bootstrap*, Tese Doutorado USP São Carlos, 2008.

[18] C. L. Balke, C. J. Merz, *UCI Repository of Machine Learning Databases. Irvine,* University of California, 1998. Acedido em 02 de março 2012, em: http://mlearn.ics.uci.edu/MLRepository.html.