Uma Estratégia Metaheurística Híbrida para Clusterização Automática e Redução da Dimensionalidade dos Dados

Éldman de Oliveira Nunes, Aura Conci
Escola de Administração do Exército, Rua Território do Amapá, 455, 41.540-830, Salvador, BA, Brasil
Universidade Federal Fluminense, Rua Passo da Pátria, 156, 24.210-240, Niterói, RJ, Brasil

***Resumo* — A clusterização é uma importante técnica para análise do comportamento dos dados e tem sido largamente utilizada para solução de diversos problemas de ordem prática. Entretanto, o desconhecimento do número ideal de grupos para partição da base e o excesso de atributos que compõe os dados contribuem para degradação da qualidade dos resultados. Este artigo apresenta uma nova estratégia para reduzir a dimensionalidade dos dados através do emprego de Algoritmos Genéticos combinados com algoritmo K-Means adaptado para clusterização automática. Os resultados demonstram ser possível melhorar a abrangência e a acurácia na classificação dos dados.**

***Palavras-chaves* — Análise de Agrupamentos, Segmentação, Algoritmos Genéticos, K-Means.**

I. INTRODUÇÃO

Análise Operacional é uma ciência que objetiva fornecer ferramentas para o auxílio ao processo de tomada de decisões por meio da construção de um modelo do problema para análise e compreensão do comportamento da situação. Tal modelo inclui variáveis de decisão, restrições e uma função objetivo que se deseja otimizar. Assim, uma solução consiste em encontrar, num espaço de busca multidimensional, um tal conjunto de valores que, ao serem aplicados às variáveis do problema, permita minimizar (ou maximizar) uma função cujas variáveis devem obedecer às restrições impostas.

O processo de tomada de decisão é efetivamente influenciado pelo grau de conhecimento da realidade existente. Muito deste conhecimento pode ser obtido a partir dos dados disponíveis. Através da análise do comportamento dos dados é possível extrair informação relevante para a identificação de perfis, a comparação de similaridades e dissimilaridades, o reconhecimento de padrões, a segmentação e classificação de grupos e a realização de previsões [1].

A clusterização é uma técnica extremamente importante para análise do comportamento dos dados e tem por objetivo organizar os objetos de uma base de dados em grupos de tal forma que os objetos dentro de um mesmo grupo sejam mais similares entre si do que com objetos de outros grupos. Entretanto, o desconhecimento do número ideal de grupos para partição da base, bem como a multidimensionalidade dos atributos que compõe os dados, são fatores que aumentam a complexidade do problema e degradam a qualidade dos resultados.

P. Autor, eldman@bol.com.br, S. Autor, aconci@ic.uff.br

A clusterização de um conjunto de dados sem o conhecimento a priori do número ideal de grupos e a redução da dimensionalidade desses dados, através da seleção dos atributos relevantes para o processo de agrupamento, são duas tarefas da classe de problemas NP-difíceis. Encontrar soluções ótimas ou aproximadas para esta classe de problemas de otimização está longe de ser uma tarefa fácil. Problemas desta complexidade são caracterizados por não linearidade, ruídos, descontinuidades ou espaço de busca extremamente grande. Normalmente, o emprego de métodos convencionais torna-se inviável em virtude do enorme esforço computacional exigido para sua solução [2].

Metaheurísticas como Algoritmos Genéticos (AG) são estratégias adequadas para solução de problemas desta natureza. Combinando escolhas aleatórias com o conhecimento obtido em resultados anteriores, um AG emprega mecanismos de busca que se guiam através do espaço de pesquisa do problema evitando paradas prematuras em ótimos locais e, conseqüentemente, proporcionando melhores soluções.

Este trabalho apresenta uma nova estratégia que combina as metaheurísticas de Algoritmos Genéticos e Algoritmo K-Means adaptado para clusterização automática, a fim de reduzir a dimensionalidade dos dados a partir da determinação automática do número ideal de grupos presentes em uma base. Os resultados obtidos com a nova estratégia proposta demonstraram ser possível melhorar a abrangência e a acurácia na classificação dos dados, o que a torna de grande utilidade para análise do comportamento dos dados e, conseqüentemente, para o processo de tomada de decisões.

I. CLUSTERIZAÇÃO

A clusterização tem por objetivo organizar os objetos de uma base de dados em grupos de tal forma que os objetos dentro de um mesmo grupo sejam mais similares entre si do que com objetos de outros grupos.

Quando a solução do problema de agrupamento depende da informação *a priori* do número de grupos desejados para partição dos dados, este é referenciado na literatura como "problema de k-clusterização" [3]. Entretanto, na maioria das aplicações práticas, o número ideal de grupos é desconhecido e sua obtenção acrescenta complexidade na solução, sendo referenciado na literatura como "problema de clusterização automática" [4].

Problemas de clusterização podem ser definidos da seguinte forma [3]:
Dado um conjunto com n elementos $X = \{X_1, X_2, ..., X_n\}$, o problema de clusterização consiste na obtenção de um conjunto de k clusters, $C = \{C_i, C_2, ..., C_k\}$, tal que haja uma maior similaridade entre os elementos contidos em um cluster C_i do que qualquer um destes com os elementos de um dos demais clusters do conjunto C.

A busca pela melhor solução no espaço de soluções possíveis torna o processo de clusterização um problema NP-Difícil. A utilização de métodos exatos para obtenção da solução ótima fica impraticável, uma vez que a verificação exaustiva de todas as configurações de agrupamentos possíveis é computacionalmente inviável [2].

Para o problema de k-clusterização, quando se conhece o número de k *a priori*, existem N diferentes maneiras de agrupar n elementos em k clusters:

$$N(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^n \quad (4)$$

O que torna exponencial o crescimento de n. Por exemplo, existem os seguintes números de soluções possíveis para se combinar 10 elementos, 100 elementos e 1000 elementos em 2 clusters, respectivamente: 512 soluções, $6{,}33825\times10^{29}$ soluções e $5{,}3575\times10^{300}$ soluções.

Para o problema de clusterização automática, o número de combinações possíveis é dado por (5):

$$N(n,k) = \sum_{k=1}^{n}\frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^n \quad (5)$$

Assim, aumenta significativamente o número de combinações possíveis. Para clusterização de um conjunto com apenas 10 elementos, em um número de clusters variável de 1 a 10, existem 115.975 maneiras diferentes.

II. MÉTODO K-MEANS

K-means é um método de partição baseado em recolocação que necessita da definição *a priori* do número de agrupamentos k. O critério de custo a ser minimizado é definido em função da distância dos elementos em relação aos centros dos agrupamentos. Usualmente, este critério é a soma residual dos quadrados das distâncias (geralmente é usada a distância euclidiana). Entende-se por soma residual dos quadrados, a soma dos quadrados das distâncias dos elementos ao centróide do seu cluster, conforme (6):

$$W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - X_i)(x_{ij} - X_i)^t \quad (6)$$

Onde x_{ij} é o j-ésimo objeto do cluster i, X_i é o representante do cluster i (a média ou mediana dos objetos do cluster), e n_i é a quantidade de objetos do cluster i.

O Quadro 1 apresenta um pseudocódigo de um K-Means:

Determinar as posições iniciais dos k centróides dos clusters;
Repita
  Alocar cada elemento ao cluster do centróide mais próximo;
  Recalcular os centros dos clusters a partir dos elementos alocados;
Até atender algum critério de convergência.

Quadro. 1. Pseudocódigo de um K-Means básico

Como critério de convergência pode se executar o algoritmo até que os centróides não se movam mais ou até que um determinado número máximo de interações seja alcançado.
A partir deste primeiro particionamento, o algoritmo realiza uma busca de um ponto de máximo para o seu critério de parada. Não há garantias de que o algoritmo encontre o máximo global, sendo possível encontrar soluções distintas em diferentes execuções do algoritmo [5].

III. CLUSTERIZAÇÃO AUTOMÁTICA

Um critério para determinar a qualidade da clusterização é apresentado em [6] e estabelece a seguinte função de k:

$$G(k) = \frac{(n-k)B}{(k-1)W} \quad (7)$$

Onde k é o número de clusters usados para a segmentação, n o número de objetos da base de dados, W conforme (5), e B é a variação não explicada do somatório dos quadrados das diferenças de toda a base de dados para a média a partir do somatório dos quadrados das diferenças de cada objeto de cada cluster para o centróide do seu cluster, dado por (8):

$$B = T - W = \sum_{i=1}^{k} n_i (X_i - X)(X_i - X)^t \quad (8)$$

T é a soma dos quadrados das diferenças de cada objeto da base de dados para a média de todos os objetos da base (X), conforme (9):

$$T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - X)(x_{ij} - X)^t \quad (9)$$

O valor de $G(k)$ representa a estatística da análise da variância do agrupamento formado. Quanto maior o seu valor, mais homogêneos serão os objetos dentro de cada grupo e melhor será a partição.

IV. ALGORITMOS GENETICOS

Algoritmos Genéticos (AG) são modelos estocásticos e probabilísticos de busca e otimização, inspirados na evolução natural e na genética [7]. Um AG atua sobre um conjunto (população) de indivíduos (cromossomos) cada um representando uma possível solução do problema. Durante o processo evolutivo os cromossomos são avaliados e selecionados para reprodução e sobrevivência. Os cromossomos selecionados para reprodução, geralmente os de maiores aptidão, poderão sofrer alterações genéticas baseadas em operadores de cruzamento e mutação gerando descendentes para próxima geração. Este processo se repete por um número determinado de gerações ou outro critério de parada [8]. Os AG's são adequados para solução de problemas de otimização complexos, dependentes de muitas variáveis e com extenso espaço de soluções. Embora nem sempre encontrem a solução ótima de otimização, são capazes de fornecer soluções próximas do ótimo, perfeitamente satisfatórias quando consideramos o grau de complexidade do problema [9]. O Quadro 2 apresenta um pseudocódigo de um AG simples.

Algorítmo Genético
Início
  Inicialize a população de cromossomos;
  Avalie indivíduos na população **P**;
  Repita (evolução)
    Repita
      Selecione 2 individuos em **P** para reprodução;
      Aplique os operadores de cruzamento e mutação;
      Insira o novo indivíduo em **P'**;
    Até população **P'** ficar completa
    Avalie indivíduos na população **P'**;
    P ← **P'**;
    Incremente a geração;
  Até o objetivo final ou número máximo de gerações
Fim

Quadro. 2. Pseudocódigo de um AG basico

V. METODOLOGIA PROPOSTA

Para determinação automática do número ideal de agrupamentos foi desenvolvido um sistema que combina o emprego de Algoritmos Genéticos e Algoritmo K-Means modificado para clusterização automática [10]. O AG é utilizado para obter um subconjunto de atributos que permita uma clusterização satisfatória dos dados de uma base de dados. Para isso usa como função de avaliação o Algoritmo K-Means modificado para particionar uma base de dados para cada um dos possíveis valores de k, maximizando (7).
Como o algoritmo K-Means possui o inconveniente de ter sua solução influenciada pela escolha da configuração inicial (eleição arbitrária dos k objetos como centros iniciais dos clusters), optou-se por inicializar os centróides de forma a dispersá-los uniformemente sobre o espaço de representação. Esta estratégia tende a fornecer melhores resultados, além de contribuir para a aceleração da convergência.

Para redução da dimensionalidade dos dados é necessário realizar, dentro do conjunto de atributos disponíveis, a seleção dos atributos mais relevantes, ou seja, que mais contribuam para separabilidade das classes. Basicamente este processo de múltiplas iterações consiste dos seguintes passos:

(1) Selecionar um subconjunto de atributos (atributos candidatos) através do emprego de AG;
(2) Realizar a clusterização automática da base de dados considerando este subconjunto de atributos;
(3) Medir a qualidade da clusterização, conforme (7);
(4) Repetir os passos 1 a 3 até que se encontre um resultado satisfatório.

VI. EXPERIMENTOS

Foram realizados dois experimentos com duas bases obtidas no repositório de bases de dados para descoberta de conhecimento da Universidade da Califórnia, Irvine (*UCI MachineLearning Repository*) [11].
Cada experimento foi dividido em duas etapas. A primeira consistiu da clusterização automática da base de dados considerando todos os seus atributos. A segunda, consistiu da clusterização automática da base de dados considerando subconjuntos de atributos obtidos com AG. A qualidade dos resultados foi avaliada através das medidas de abrangência (10) e acurácia (11), bem como da matriz de classificação, que indica o número de amostras classificadas em cada classe. A situação ideal é estar toda a amostragem de uma classe na diagonal principal. Valores fora da diagonal principal indicam erro de classificação.

$$Abrangência = \frac{Número\ de\ Amostras\ da\ Classe\ no\ Cluster}{Número\ Total\ de\ Amostras\ da\ Classe} \quad (10)$$

$$Acurácia = \frac{Número\ de\ Amostras\ da\ Classe\ no\ Cluster}{Número\ Total\ de\ Amostras\ do\ Cluster} \quad (11)$$

A. Primeiro caso de teste

Para realização do primeiro caso de teste, foi utilizada a popular base de dados *Iris Plants*. A base Íris consiste de 3 classes: Íris-Setosa, Íris-Virginica e Íris-Versicolor, cada uma possuindo 50 instâncias, com distribuição de 33.3% para cada uma das 3 classes. Cada instância possui cinco atributos, sendo quatro do tipo numérico (comprimento da sépala em cm, largura da sépala em cm, comprimento da pétala em cm, largura da pétala em cm) e um do tipo categórico (classe: Íris-Setosa, Íris-Versicolor, Íris-Virginica). O atributo do tipo categórico não participa do processo de clusterização, entretanto define a qual classe a instância pertence.

1) Primeira etapa: clusterização automática com utilização de todos os atributos

A Fig. 1. apresenta a clusterização da base Íris considerando suas 150 instâncias, cada uma com seus quatro atributos. Para este teste, o valor de k variou entre 2 e 150.

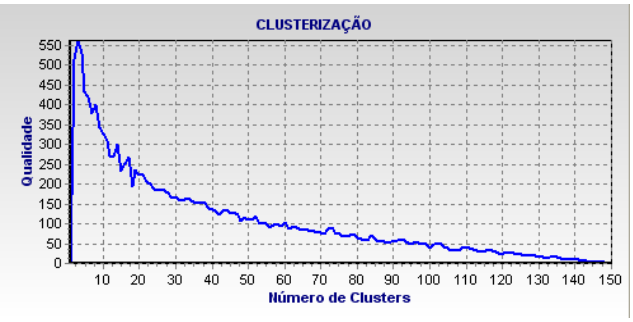

Fig. 1. Clusterização com k variando entre 2 e 150

A Fig. 2. apresenta a mesma clusterização com o valor de k variando entre 2 e 10. É possível constatar que o número de clusters encontrado corresponde ao número de classes existentes na base, ou seja, igual a três.

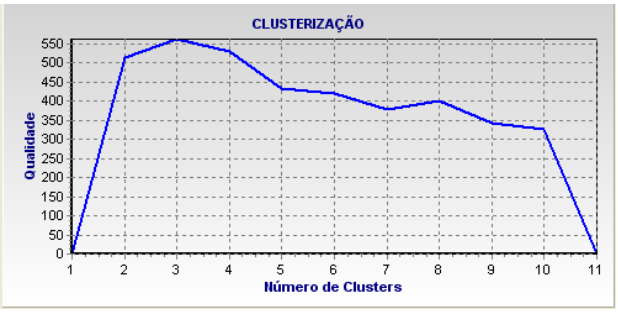

Fig. 2. Clusterização com k variando entre 2 e 10

A Tabela 1 apresenta os resultados obtidos com os respectivos valores de abrangência e acurácia, bem como a matriz de classificação correspondente. De sua análise verifica-se que a classe Íris-Setosa é linermente separável, enquanto que as classes Íris-Vesicolor e Íris-Verginica não.

TABELA 1 - RESULTADO COM TODOS ATRIBUTOS

Cluster	Classe	Abrangência	Acurácia
1	Iris-setosa	100,00%	100,00%
2	Iris-versicolor	94,00%	77,05%
3	Iris-virginica	72,00%	92,31%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
1	Iris-setosa	50	0	0	50
2	Iris-versicolor	0	47	14	61
3	Iris-virginica	0	3	36	39
TOTAL		50	50	50	150

2) Segunda etapa: seleção dos atributos com AG

A segunda etapa de testes consistiu do emprego do AG para determinação de um subconjunto de atributos satisfatório para clusterização da base. A Tabela 2 apresenta o melhor resultado obtido: a remoção do atributo comprimento da sépala. A clusterização foi realizada com o valor de k variando entre 2 e 10.
Pode-se verificar na Tabela 1 e na Tabela 2 que a abrangência da classe Íris-Vesicolor se manteve constante em 94% e acurácia melhorou de 77,05% para 92,16%, enquanto na classe Íris-Virginica a abrangência melhorou de 72% para 92% e a acurácia de 92,31% para 93,88%. O que indica que o conjunto de atributos (largura da sépala, comprimento da pétala e largura da pétala) é relevante para separabilidade das classes permitindo obter um resultado melhor com três atributos do que com quatro, o que representa uma redução de 25% na dimensionalidade dos dados.

TABELA 2 - RESULTADO COM SUBCONJUNTO DE ATRIBUTOS

Cluster	Classe	Abrangência	Acurácia
1	Iris-setosa	100,00%	100,00%
2	Iris-versicolor	94,00%	92,16%
3	Iris-virginica	92,00%	93,88%

Cluster	Classe	Iris-setosa	Iris-versicolor	Iris-virginica	TOTAL
1	Iris-setosa	50	0	0	50
2	Iris-versicolor	0	47	4	51
3	Iris-virginica	0	3	46	49
TOTAL		50	50	50	150

B. Segundo caso de teste

Para realização do segundo caso de teste, foi utilizada a base de dados Heart. Esta base consiste de 2 classes: ausência ou presença de doença do coração, sendo 150 amostras de presença (55.56%) e 120 de ausência (44.44%), totalizando 270 instâncias na base. Cada instância possui quatorze atributos, sendo treze do tipo numérico (idade; sexo; tipo de dor do torax [4 valores]; pressão sanguínea em repouso; colesterol do soro em mg/dl; açúcar no sangue em jejum >120 mg/dl; resultados eletrocardiográficos em repouso [valores 0,1,2]; frequência cardíaca máxima alcançada; angina induzida em exercício; pico da idade = depressão ST induzida pelo exercício, em relação ao descanso; a inclinação do segmento ST do exercício de pico, número de veias principais [0-3] coloridas por flourosopy; thal: 3 = normal, 6 = defeito estabelecido, 7 = defeito reversível) e um do tipo categórico (classe: ausência, presença). O atributo do tipo categórico não participa do processo de clusterização, entretanto define a qual classe a instância pertence.

1) Primeira etapa: clusterização automática com utilização de todos os atributos

A Fig. 3. apresenta a clusterização da base Heart considerando suas 270 instâncias, cada uma com seus treze atributos. Para este teste, o valor de k variou entre 2 e 270.

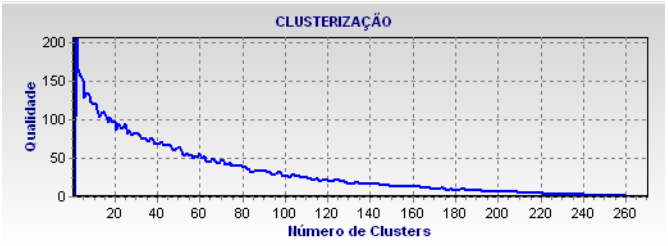

Fig. 3. Clusterização com k variando entre 2 e 270

A Fig. 4. apresenta a mesma clusterização com o valor de k variando entre 2 e 10. É possível constatar que o número de clusters encontrado corresponde ao número de classes existentes na base, ou seja, igual a dois.

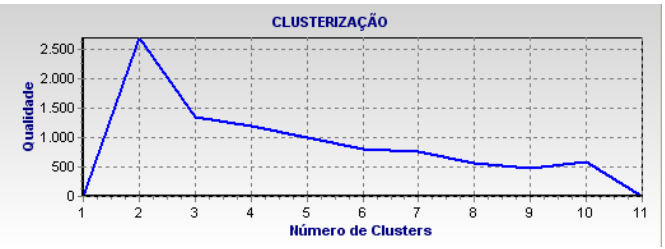

Fig. 4. Clusterização com k variando entre 2 e 10

A Tabela 3 apresenta os resultados obtidos com os respectivos valores de abrangência e acurácia, bem como a matriz de classificação correspondente. De sua análise verifica-se que as classes não são linearmente separáveis.

TABELA 3 - RESULTADO COM TODOS ATRIBUTOS

Cluster	Classe	Abrangência	Acurácia
1	Ausência de doença do coração	69,33%	61,90%
2	Presença de doença do coração	46,67%	54,90%

Cluster	Classe	Ausência de doença do coração	Presença de doença do coração	TOTAL
1	Ausência de doença do coração	104	64	168
2	Presença de doença do coração	46	56	102
TOTAL		150	120	270

2) Segunda etapa: seleção dos atributos com AG

A segunda etapa de testes consistiu do emprego do AG para determinação de um subconjunto de atributos satisfatório para clusterização da base. A Tabela 4 apresenta o melhor resultado obtido. Foram removidos dez atributos e mantidos apenas três atributos: açúcar no sangue em jejum> 120 mg/dl, angina induzida em exercício e thal. A clusterização foi realizada com o valor de k variando entre 2 e 10.

TABELA 4 - RESULTADO COM SUBCONJUNTO DE ATRIBUTOS

Cluster	Classe	Abrangência	Acurácia
1	Ausência de doença do coração	79,33%	78,29%
2	Presença de doença do coração	72,50%	73,73%

Cluster	Classe	Ausência de doença do coração	Presença de doença do coração	TOTAL
1	Ausência de doença do coração	119	33	152
2	Presença de doença do coração	31	87	118
TOTAL		150	120	270

Pode-se verificar na Tabela 3 e na Tabela 4 que a abrangência da classe "Ausência de doença do coração" melhorou de 69,33% para 79,33% e a acurácia melhorou de 61,90% para 78,29%, enquanto na classe "Presença de doença do coração" a abrangência melhorou de 46,67% para 72,50% e a acurácia de 54,90% para 73,73%. O que indica que o subconjunto de atributos selecionados é relevante para separabilidade das classes permitindo obter um resultado melhor com três atributos do que com treze, o que representa uma redução de 76,92% na dimensionalidade dos dados.

VII. CONCLUSÃO

Este trabalho combinou as seguintes estratégias para solução do problema de clusterização automática e da redução da dimensionalidade dos dados: implementação do Algoritmo Genético para seleção de atributos candidatos relevantes, implementação do método K-Means modificado para clusterização automática, inicialização dos centróides com dispersão uniforme sobre o espaço de representação e maximização do valor da função $G(k)$.

A fim de avaliar os resultados, realizou-se a clusterização da base de dados Íris e Heart. Para ambas as bases de dados foi possível melhorar a abrangência e a acurácia da clusterização, além de reduzir a dimensionalidade do dados em 25% para base Íris e 76,92% para base Heart.

Os resultados preliminares indicam que a metodologia proposta é capaz de determinar automaticamente o número ideal de clusters de uma base, bem como reduzir a dimensionalidade dos dados através da seleção dos atributos mais relevantes para separabilidade das classes. Tais funções são de grande utilidade para análise do comportamento dos dados.

Para trabalhos futuros utilizar-se-á a metodologia proposta em aplicações de ordem prática em diversas áreas de interesse.

REFERÊNCIAS

[1] HAN, J.; KAMBER, M. Data Mining Concepts and Techniques. San Francisco: Morgan Kaufmann, 2001. 550 p.

[2] OCHI, L.S. Problemas de Clusterização em Mineração de Dados. In: Encontro Regional de Informática RJ/ES, 2004, Vitória. Anais do ERI 2004 RJ/ES, Vitória: ERI, 2004. p.1-6. CD-ROM.

[3] FASULO, D. An analysis of recent work on clustering algorithms. Technical report, 01-03-02, Seattle: University Of Washington - Department Of Computer Science And Engineering, 1999. 23 p.

[4] DOVAL, D.; MANCORIDIS, S.; MITCHELL, B.S. Automatic Clustering of Software Systems Using a Genetic Algorithm. step, p. 73, Software Technology and Engineering Practice, 1999.

[5] CAMASTRA, F.; VERRI, A. A Novel Kernel Method for Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 27, n. 5, 2005.

[6] CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods, v. 3, (1), p. 1-27, 1974.

[7] HOLLAND, J. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, 1975.

[8] Goldberg, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley Publishing Company, Inc., 1989.

[9] Linden, Ricardo. Algoritmos Genéticos: Uma importante ferramenta da Inteligência Computacional. Rio de Janeiro: Brasport, 2006. 348p.

[10] NUNES, Eldman de Oliveira; CONCI, A. Clusterização Automática na Redução da Dimensionalidade dos Dados. In: XI Simpósio de Pesquisa Operacional e Logística da Marinha, 2008, Rio de Janeiro. XI Simpósio de Pesquisa Operacional e Logística da Marinha. Rio de Janeiro: Centro de Análises de Sistemas Navais - Secretaria de Ciência, Tecnologia e Inovação da Marinha, 2008. p. 1-14.

[11] ASUNCION, A.; NEWMAN, D.J. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 2007.