CLUSTERIZAÇÃO COM ALGORITMO GENÉTICO BASEADO EM DENSIDADE E GRADE

Fábio dos S. Ferreira
Mestrado Acadêmico em Ciência da Computação (MACC), Universidade Estadual do Ceará (UECE)
Av. Paranjana, 1700 – 60.740-903 – Fortaleza, CE – Brasil
fabio@mec2a.org

Gustavo A. L. de Campos
Mestrado Acadêmico em Ciência da Computação (MACC), Universidade Estadual do Ceará (UECE)
Av. Paranjana, 1700 – 60.740-903 – Fortaleza, CE – Brasil
gustavo@larces.uece.br

Jackson S. de V. Silva
Mestrado Acadêmico em Ciência da Computação (MACC), Universidade Estadual do Ceará (UECE)
Av. Paranjana, 1700 – 60.740-903 – Fortaleza, CE – Brasil
jackson@larces.uece.br

RESUMO

O Algoritmo Genético é amplamente utilizado na tarefa de análise de agrupamento, mas a sua utilização normalmente é encontrada na literatura baseada na metodologia de densidade para a formação dos grupos. A metodologia de grade traz como benefício à constituição de grupos de formatos arbitrários, com um menor tempo computacional. Este artigo apresenta um novo Algoritmo Genético para clusterização baseado nas metodologias de densidade e grade, utilizando para a formação dos grupos as células formadas pela adição da grade. Através do benefício da busca global do Algoritmo Genético juntamente com a sua facilidade de implementação, o objetivo deste trabalho é encontrar os melhores centros dos grupos de forma mais rápida. Graças à união dos métodos de densidade e grade ocorre a redução da complexidade do problema, tornando a busca pela solução menos custosa.

PALAVRAS CHAVE. Clusterização, Algoritmo Genético, Densidade, Grade.

ABSTRACT

The Genetic Algorithm is widely used in the task of cluster analysis, but its use is usually found in the literature based on the methodology of density for the formation of clusters. The grid methodology brings benefits to formation of clusters of arbitrary formats with a smaller computational time. This paper presents a new genetic algorithm for clustering based on the methodologies of density and grid, using the formation of cell clusters formed by the addition of the grid. Through the benefit of the global search of Genetic Algorithm together with its facility of implementation, the objective is to find the best centers of the groups more quickly. Thanks to the union of the methods density and grid occurs the reduction the complexity of the problem, making the search for less costly solution.

KEYWORDS. Clustering, Genetic Algorithms, Density, Grid.

1. Introdução

O Algoritmo Genético (GA – *Genetic Algorithm*) é reconhecido como uma poderosa ferramenta para solucionar problemas de otimização (Lorena & Furtado, 2001). O fundamento deste algoritmo é a evolução controlada de uma estrutura populacional.

Segundo (Oliveira, 2007) os Algoritmos Evolutivos (AE), entre estes o GA, "são caracterizados por realizarem uma busca globalizada, paralela e otimizada em direção ao ótimo global de uma função. Estas características propiciam aos AEs obterem um ganho na qualidade das soluções obtidas". Esta característica o torna um alvo de estudo em diversos campos que possuem tarefas computacionalmente difíceis.

Segundo (Oliveira, 2007) a análise de agrupamento pode ser definida como o problema de agrupar n objetos em m grupos. Considerando $n \geq m$, à medida que o tamanho de n e m aumentam, o conjunto de todas as combinações de grupos possíveis se torna muito grande, o que o faz pertencer à classe de problemas referenciados como NP-difíceis, onde o "tempo de execução para um algoritmo conhecido qualquer para garantir uma solução ótima é uma função exponencial do tamanho do problema" (Eglese, 1990).

A análise de agrupamento é o estudo formal de algoritmos e métodos para agrupar objetos, e tem o objetivo de encontrar uma organização para dados muitas vezes não explorados (Oliveira, 2007).

Conhecida na biologia como "taxonomia numérica", compreende processos interdependentes para classificar objetos descritos como um conjunto de valores de diversas variáveis (Spath, 1980). A aplicação de métodos de clusterização é apropriada se mais do que duas variáveis devem ser consideradas ao mesmo tempo. Segundo Spath valores realísticos para o numero de objetos está em torno de 50 à 5000 e suas variáveis em torno de 1 à 10.

As técnicas de clusterização agrupam um conjunto de dados em um espaço de d dimensões para maximizar a similaridade entre os grupos e minimizar a similaridade entre dois grupos diferentes (Garai & Chaudhuri, 2004).

Para uma boa implementação de algoritmos de agrupamento é desejável que ele cumpra alguns requisitos: a) Escalabilidade, o algoritmo deve funcionar bem em bases de dados de qualquer tamanho, grande ou pequenas; b) habilidade para lidar com diferentes tipos de atributos sejam eles numéricos, categóricos, reais, entre outros; c) descoberta de grupos de formatos variados, alguns algoritmos tendem a encontrar grupos esféricos com tamanho e densidade similares, é desejável que o algoritmo possa encontrar grupos com formatos arbitrários; d) habilidade para lidar com dados contendo ruídos, o algoritmo deve conseguir lidar com bases com valores discrepantes, dados desconhecidos, errados, ou até mesmo a falta de dados; e) habilidade de lidar com bases multidimensionais (Oliveira, 2007).

Há uma grande quantidade de algoritmos de agrupamentos, o que torna difícil a sua classificação. Em Han & Kamber, (2006) é encontrado uma forma de classificá-los de acordo com o método utilizado para a formação dos grupos. A seguir será apresentado dois destes métodos que serviram como base para a formulação do algoritmo proposto, o método de densidade e o método de grade.

1.1. Método de Densidade

Os métodos baseados em densidade permitem descobrir grupos de formatos arbitrários. Estes métodos consideram grupos como sendo regiões densas de objetos no espaço de dados que são separados por regiões de baixa densidade, que geralmente representam ruídos (Han & Kamber,

2006). Um exemplo de algoritmo baseado em densidade é o algoritmo DBSCAN (Ester et. al., 1996).

O algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) cria grupos de objetos em regiões de alta densidade e é capaz de descobrir grupos de formatos arbitrários em bases de dados espaciais e com ruídos. Para o DBSCAN um grupo é definido como um conjunto máximo de pontos conectados por densidade (Oliveira, 2007).

1.2. Método de Grade

Os métodos baseados em grade dividem os objetos em um número finito de células que formam uma estrutura de grade multidimensional na qual todas as operações de agrupamento são realizadas. A principal vantagem do método é que as operações independem do número de objetos da base de dados, e sim do número de células da estrutura da grade, o que melhora a desempenho dos algoritmos baseados nesta heurística (Han & Kamber, 2006). Um exemplo de algoritmo baseado em grade é o algoritmo CLIQUE (Rakesh et. al., 1999).

O algoritmo CLIQUE (*Clustering in Quest*), baseado em grade e em densidade, particiona o conjunto de dados em subespaços (grade de células) para encontrar agrupamentos suficientemente densos. Cada grade organiza um conjunto de dados, separando os valores contínuos de cada atributo em um número de intervalos discretos. Por fim, cada objeto é atribuído a uma célula a qual seu intervalo contém o valor original do objeto. Os agrupamentos são formados a partir da junção de células densas adjacentes (Oliveira, 2007).

No tópico 2 será apresentado uma relação de trabalhos relacionados ao tema apresentado aqui. O tópico 3 fará uma breve apresentação de algoritmo genético e seus conceitos fundamentais. No tópico 4 será realizada a apresentação do algoritmo resultante deste trabalho, também será mostrado como os operadores genéticos do GA foram utilizados, e uma descrição dos processos necessários para utilizá-lo. O tópico 5 apresenta os resultados referentes aos testes realizados com o algoritmo em bases de dados públicas. No tópico 6 uma conclusão a cerca do trabalho como um todo.

2. Trabalhos Relacionados

Em Maulik & Bandyopadhyay (2000), o algoritmo de clusterização *k-means* (Macqueen, 1967) é aplicado no Algoritmo Genético graças a sua simplicidade. Na fase de avaliação do cromossomo, o calculo do *fitness* no GA, é utilizado o calculo das distancias entre os centros, codificados no cromossomo, e cada ponto da base. Caso a distancia entre um ponto e um centro seja menor do que a distancia do ponto para qualquer outro centro, então o ponto é adicionado a este grupo de menor distancia, assim como no *k-means*.

Em Mota & Gomide (2005) é aplicada a técnica de *fuzzy c-means* de Hall et al. (1999) na fase de avaliação do cromossomo, no GA. Para que com isso seja possível reduzir a avaliação direta dos objetos, e assim agilizar o processo. O trabalho se baseia na idéia de que um objeto pode pertencer a mais de um grupo.

Em Garai & Chaudhuri (2004) os dados são decompostos em alguns grupos fragmentados. O Algoritmo Genético é aplicado sobre esses fragmentos a fim de procurar a melhor formação dos grupos. Esses grupos encontrados serão submetidos a uma fase de união, onde serão analisados pela sua adjacência de forma que estes possam ser unidos em um novo grupo.

No trabalho de Oliveira (2007) é criado um novo Algoritmo Evolutivo para a tarefa de análise de agrupamento chamado EDACluster, baseado no algoritmo EDA (Algoritmo de Estimativa de Distribuição - *Estimation of Distribution Algorithms*) o qual não utiliza os operadores de

cruzamento e mutação, mas sim uma amostragem da distribuição de probabilidade da população. Em seu trabalho aplica uma metodologia hibrida para formação de grupos baseada nos métodos de densidade e grade.

3. Algoritmo Genético

Os Algoritmos Genéticos foram inventados por John Holland em 1960, com o objetivo de abstrair os processos adaptativos dos sistemas naturais, e com isso desenvolver softwares de sistemas artificiais, capazes de simular os mecanismos dos sistemas naturais (Teixeira, 2005). Pertencem a uma família de modelos computacionais, inspirados pela evolução. Esses algoritmos codificam uma solução em potencial para um problema específico em alguns cromossomos simples, em forma de estrutura de dados, aplicando operadores de recombinação nessas estruturas para assim preservar informações críticas (Whitley, 1994).

Os seus paradigmas consistem de buscas estocásticas que são baseadas na abstração do processo de evolução Neo-Darwiniano. A idéia básica é que um "indivíduo" da população envolvida codifica uma solução candidata para dado problema. Cada indivíduo é avaliado por uma função de *fitness*. Esses indivíduos evoluem através da relação entre os melhores indivíduos através dos operadores baseados na seleção natural, i.e. sobrevivência e reprodução do mais apto, e genética, onde são aplicados os operados de cruzamento e mutação (Freitas, 2002).

O operador de cruzamento essencialmente armazena material genético entre dois indivíduos. O operador de mutação modifica o valor de um gene para um novo valor randômico. Ambos, cruzamento e mutação, são operadores estocásticos, aplicados com probabilidades definidas pelo usuário. A probabilidade de mutação é usualmente menor que a de cruzamento (Freitas, 2002).

4. Clusterização com Algoritmo Genético Baseado em Densidade e Grade

Este artigo se baseia no trabalho se Cesar Siqueira (Oliveira, 2007), no qual criou um algoritmo evolutivo baseado em densidade e grade para análise de agrupamento, chamado EDACluster. Em seu trabalho o GA utilizado não faz uso dos operadores de cruzamento e mutação, mas se baseia no algoritmo EDA (Larrañaga & Lozano, 2002). O EDA substitui os operadores de mutação e cruzamento tradicionais pelo aprendizado e amostragem da distribuição de probabilidade dos melhores indivíduos da população a cada interação do algoritmo (Oliveira, 2007).

O algoritmo EDACluster utiliza o algoritmo EDA aplicando-o em análise de agrupamento, utilizando como método de formação de grupos uma forma hibrida dos métodos de densidade e grade. O algoritmo busca identificar áreas densas na base de dados, definidas por conjuntos de células densas adjacentes. O algoritmo identifica os *k* centros dos grupos e então as células densas mais próximas serão adicionadas ao grupo, desde que estas não diminuam sua densidade abaixo de um limiar (Oliveira, 2007).

Este artigo apresentará um GA aplicado em análise de agrupamento, que também utiliza o método híbrido baseado em densidade e grade, mas utilizando a forma mais comum do Algoritmo Genético com cruzamento e mutação. Para a formação dos grupos é utilizada a distância mínima que separa uma célula densa aos centros dos grupos, ao invés de um limiar de redução de densidade.

O GA utiliza um método de seleção dos indivíduos que participarão dos operadores genéticos de cruzamento proporcional ao seu *fitness* (adaptabilidade do cromossomo). Conforme Whitley (1994), a operação de seleção escolhe cromossomos da população para a reprodução utilizando na maioria das vezes a adaptabilidade do cromossomo, de modo que os indivíduos mais

aptos sejam selecionados mais vezes, da mesma forma como ocorre nos mecanismos da seleção natural.

Existem diversos métodos se seleção, mas para este trabalho foi utilizado o método da roleta, onde "a probabilidade de que certo indivíduo seja selecionado é igual à sua aptidão (*fitness)* dividida pela média da aptidão de toda a população" (Teixeira, 2005). Segundo afirma Mitchell (1999), isto é equivalente ao que os biologistas chamam de "Viabilidade de Seleção".

O cruzamento é um processo de recombinação de partes das seqüências de caracteres entre pares de cromossomos, selecionados como pais, com o objetivo de gerar nova descendência, sendo esta composta de indivíduos melhores adaptados, ou seja, com valores maiores de *fitness* (Fogel, 2000).

O método utilizado para "cruzar" os cromossomos foi o cruzamento com 1 (um) ponto de corte. Os Algoritmos Genéticos tradicionalmente utilizam este tipo de operador de cruzamento, onde dois indivíduos da população, após a seleção, são submetidos ao processo de cruzamento, que se dá da seguinte maneira: o ponto de corte é aleatoriamente gerado, e este é menor ou igual ao tamanho do cromossomo. Depois, os caracteres que precedem o ponto de corte são preservados, e os caracteres posteriores são trocados entre o par (pais) participante do processo (Teixeira, 2005). Foi escolhido este método principalmente por que o cromossomo utilizado normalmente possui tamanho pequeno, de 3 (três) a 5 (cinco) grupos (genes).

De acordo com Fang (1994), o operador de mutação tem o poder de garantir que todos os cromossomos possíveis sejam alcançados. A operação de mutação deve ser aplicada a uma pequena porcentagem da população em cada geração, acarretando em uma baixa taxa de mutação (Teixeira, 2005). Caso seja aplicada uma taxa mais elevada o Algoritmo Genético passará a se comportar como uma busca aleatória (Mitchell, 1999).

A mutação utilizada neste trabalho modifica o índice de um gene do cromossomo, centro, pelo índice de uma célula que não esteja contido neste cromossomo, aleatoriamente. Garantindo que o maior número possível de células seja coberto pelo algoritmo. Buscando assim o melhor conjunto de células à serem utilizadas como centros para os grupos. Assim cobrindo o maior numero possível de células, conseqüentemente os objetos da base.

A seguir será apresentada a forma como este algoritmo foi modificado para ser aplicado em análise de agrupamento.

4.1. Configuração

Para a utilização do algoritmo primeiramente é necessário que a base de dados a ser utilizada, passe por uma etapa de configuração. Nesta etapa a base passará a ser representada por uma grade de *d* dimensões. Cada dimensão será representada por um atributo dos dados.

Cada atributo deverá ser dividido em um numero de faixas para os seus valores. Estes deverão ser igualmente espaçados, dessa forma os itens passarão a ser representados em células que formam a grade multidimensional. Com isso, como as células possuem o mesmo volume, a densidade é definida pelo número de itens contidos em uma célula (Oliveira, 2007).

Nesta etapa os objetos não serão mais representados por seus valores de atributos, mais sim pelas células que irão ocupar na grade.

Para a base poder ser utilizada neste algoritmo as células que não possuem objetos deverão ser descartadas. Com isso caso o numero faixas seja muito grande, no pior caso o número de células que será utilizada será igual ao numero de objetos da base. O que torna o fator número de faixas e o intervalo em que os atributos serão divididos importantes no momento de utilizar o algoritmo.

A figura 1 apresenta um exemplo da etapa de configuração de uma base de dados bidimensional. Os atributos Salário e Idade foram divididos em 6 (seis) faixas de valores. Os itens

passarão a ser representados pelas células em que está contido. Como é mostrado para os objetos que estão entre os valores 30-35 do atributo idade e 300-600 para o atributo salário. Com isto eles passarão a ser representados pela célula (1,3) cuja densidade é igual a 2 (dois).

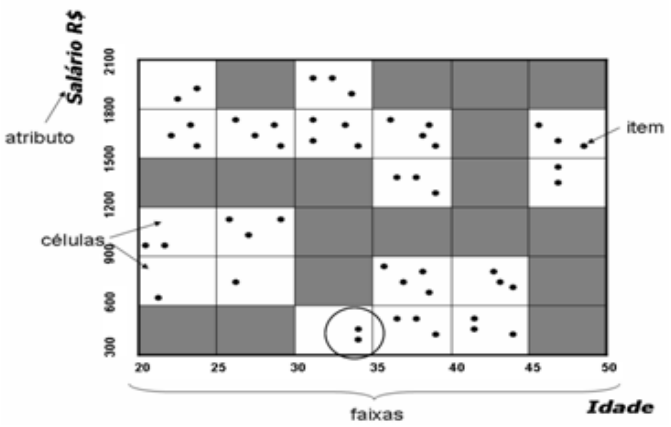

Figura 1 - Representação do plano da base em células de uma grade. Fonte: Oliveira (2007)

4.2. Representação do Indivíduo

Existem diversas formas de representar um indivíduo em uma aplicação de algoritmos genéticos. A maioria das aplicações de GA utiliza seqüências de bits (Bit String) de tamanho e ordem fixos para representar as soluções candidatas (Teixeira, 2005). Mas para este trabalho foi utilizada a formulação de Código de Permutação.

Este tipo de codificação, também conhecida como Representação por Caminho, é comumente usado em problemas de disposição e arrumação, como o Problema do Caixeiro Viajante (PCV), e em problemas de ordenação de tarefas (Teixeira, 2005).

Neste trabalho cada indivíduo do GA será representado por um vetor (cromossomo) e o tamanho deste vetor se refere ao número de grupos que o usuário deseja que o algoritmo encontre. Cada posição (gene) deste vetor conterá um índice, 1, 2, 3...*n*, referente a uma célula densa da grade.

A figura 3 apresenta um exemplo de um indivíduo (cromossomo) do GA. Cada posição representa um centro, o cromossomo pode ter *k* centros de acordo com o desejado pelo usuário. Os *n* centros podem ser inseridos em qualquer posição no cromossomo, mas o mesmo centro não pode se repetir no mesmo cromossomo.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Figura 2 - Representação de uma base de dados bidimensional, em que cada célula está representada por um índice. As células em destaque representam os centros escolhidos no indivíduo da figura 3.

CENTRO 1	CENTRO 2	CENTRO 3	CENTRO 4
2	8	9	15

Figura 3 - Representação de um indivíduo com 4 (quatro) genes (centros).

4.3. Formando o Grupo

Para a formação dos grupos é utilizado um método que verifica as distancias das células densas da grade a cada um dos centros escolhidos (genes do cromossomo). Caso uma célula *i* esteja mais próxima de um cento *j* então esta célula é atribuída ao grupo K_j correspondente a este centro, ou seja:

$$\|x_i - z_j\| < \|x_i - z_p\|, \text{ p=1,2,...,}k,\ j \neq p.$$

A distância Spearman Footrule (Teknomo, 2010) é utilizada para calcular a distância entre as células ocupadas da grade multidimensional e os centros candidatos dos grupos.

$$s_{ij} = \sum_{a=1}^{A} |x_{ia} - z_{ja}|$$

onde:
s_{ij}: distancia da célula i para o centro j;
A: número de dimensões (atributos) da grade;
a: índice do atributo para célula atual;
x_{ia}: posição da célula *i* no atributo a;
z_{ja}: posição da célula-centro *j* no atributo a;

Como exemplo pode ser utilizado o cromossomo da figura 3. A distancia da célula 1 para a célula-centro 9:

$$s_{ij} = |1\text{-}3|+|1\text{-}1| \therefore s_{ij} = 2+0 \therefore s_{ij} = 2$$

A distancia da célula 1 para a célula-centro 9 é igual a 2 (dois). Então esta célula é adicionada e a densidade atual é atualizada com a densidade da célula e a próxima célula será avaliada.

4.4. Função de Avaliação Utilizada

Após os grupos serem formados, é possível avaliar os grupos gerados baseados nas suas densidades. A função de avaliação é inspirada no algoritmo de agrupamento baseado em densidade CLIQUE (Agrawal et al., 1998). A densidade é calculada pela razão entre o número de itens de um grupo e o número de células do grupo. Após o calculo da densidade, realizado no momento da formação dos grupos, é possível calcular a função de avaliação de um indivíduo que corresponde à média das densidades de cada grupo formado (Oliveira, 2007).

$$Densidade(grupo) = \frac{Número\ de\ ítens}{Número\ de\ células}$$

$$Fitness_i = \frac{\sum_{j=1}^{k} Densidade(j)}{k}$$

5. Resultados

Os testes a seguir foram feitos em um computador Intel® CoreTM2 Duo T5800 2.00GHz, com 2GB de memória RAM, sistema operacional Microsoft Windows XP x64.

As bases de dados utilizadas foram a Iris e Auto-mpg (Newman, 2010). A base Iris possui 150 objetos classificados em 3 grupos já definidos. A base Auto-mpg possui 392 objetos também classificados em 3 grupos já definidos originalmente.

Para o GA foram utilizados 50 indivíduos, com um numero diferente de gerações para cada execução com taxa de cruzamento de 80% e taxa de mutação de 10%. Para a etapa de configuração foram utilizadas 10 faixas aplicadas em cada uma das bases em teste.

O algoritmo *k-means* utilizado é o encontrado no framework de Mineração de Dados WEKA (University of Waikato, 2010).

A tabela 1 apresenta os resultados para o teste realizado com a base de dados auto-mpg e mostra uma comparação entre os resultados encontrados com o algoritmo apresentado neste trabalho, o algoritmo de agrupamento *k-means* e o agrupamento originalmente definido na base. É possível notar que os três agrupamentos são diferentes, no entanto o GA apresentou resultados mais parecidos com o agrupamento original do que o próprio *k-means*.

O teste com o GA utilizou um total de 200 iterações com um tempo total de 161 segundos (2min 41seg), enquanto que o *k-means* utilizou um total de 200 iterações com um tempo total de menos de 1 segundo.

Tabela 1. Resultados encontrados no teste com a base de dados Auto-mpg.

	GA	%	*k-means*	%	Original	%
1	287	73.22	192	49	244	62
2	30	7.65	98	25	69	18
3	75	19.13	102	26	79	20

A tabela 2 apresenta os resultados para os testes realizados para a tabela Iris e também mostra uma comparação do algoritmo GA com o algoritmo *k-means* e o agrupamento original. É notado que neste caso o algoritmo *k-means* encontrou resultados mais parecidos com o agrupamento original.

Tabela 2. Resultados encontrados no teste com a base de dados Iris.

	GA	%	*k-means*	%	Original	%
1	97	64.67	61	41	50	33,33
2	42	28	50	33	50	33,33
3	11	7.33	39	26	50	33,33

O teste com o GA utilizou um total de 100 iterações com um tempo total de 16 segundos, enquanto que o *k-means* utilizou um total de 100 iterações com um tempo total de menos de 1 segundo.

6. Conclusão

Durante os estudos realizados para este trabalho notou-se que existem poucos, ou nenhum, trabalho utilizando o Algoritmo Genético com a metodologia de grade, apesar de este ser bastante utilizado em clusterização e ser considerado de grande utilidade na área.

Como pode ser observado nas tabelas, o resultado do teste realizado com a base Auto-mpg é considerado melhor do que os resultados encontrados para o teste com a base Iris. Um dos fatores pode ser o fato de que a divisão da base em grade favoreceu a reunião de mais objetos em uma única célula para a base Iris, já que a quantidade de dados desta é menor, e talvez seus valores mais próximos.

Mesmo com essa observação a técnica é considerada útil e seus resultados são considerados satisfatórios e melhorias podem ser feitas, tanto na etapa de formação dos grupos, assim como, também, na etapa de avaliação. Mas com estes resultados o algoritmo pode ser aplicado em trabalhos futuros para bases de dados reais, para validar os seus resultados.

Estudos podem ser realizados sobre os parâmetros utilizados, tanto os utilizados pelo Algoritmo Genético quanto os utilizados na configuração da base de dados.

Este trabalho tem objetivo de se tornar mais uma referencia sobre o assunto abordado, e contribui com uma metodologia nova para Algoritmo Genético aplicado em clusterização, que é a metodologia híbrida de densidade e grade.

7. Agradecimentos

Fábio Ferreira é apoiado pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e é um pesquisador do LADE (Laboratório de Apoio à Decisão Especialista).

Referências

GArawal, R., Gehrke, J., Gunopulos, D. & RGAhavan, P. (1998), Automatic subspace clustering of high dimensional data for data mining applications, *In ACM-SIGMOD Int. Conf. ManGAement of Data*, Washington. p. 94-105.

Eglese, R. (1990), Simulated Annealing - A Tool of Operational Research, *European Journal of Operational Research*, Vol. 46, pp. 271-281.

Fang, H. L. (1994), Genetic Algorithms in Timetabling and Scheduling, *PhD Thesis*, Department of Artificial Intelligence, University of Edinburgh.

Fogel, D. B. (2000), Evolutionary Computation: toward a new philosophy of machine intelligence, *Second Edition*, IEEE Press.

Garai, G. & Chaudhuri, B. (2004), A novel genetic algorithm for automatic clustering, *Pattern Recognition Letters*, Ed. 25, pg. 173–187.

Hall, L.O., Ozyurt, I.B., Bezdek, J.C. (1999), Clustering with a genetically optimized approach, *IEEE Trans. Evolut. Comput*, 3 (2), 103–112.

Han, J. & Kamber, M. (2006). Data Mining: Concepts and Techniques - Second Edition. *Elsevier*.

Larranãga, P. & Lozano, J. (2002), Estimation of Distribution Algorithms: A new tool for Evolutionary Computation, *Kluwer Academic Publishers*, p. 382.

Lorena, L. & Furtado, J. (2001), Constructive Genetic Algorithm for Clustering Problems, *Massachusetts Institute of Technology*, Evolutionary Computation, v. 9, n. 3.

Maulik, U. & Bandyopadhyay, S. (2000), Genetic Algorithm-based Clustering Technique, *Pattern Recognition* p.33,1455-1465.

Mota, F. & Gomide, F. (2005), Hybrid Genetic Algorithms and Clustering. *State University of Campinas*, Campinas – SP, Brazil.

Newman, D., Hettich, S., Blake, C. & Merz, C., UCI Repository of machine learning databases, *University of California*, Department of Information and Computer Science, http://www.ics.uci.edu/~mlearn/MLRepository.html, Acesso em 29 mar 2010.

Oliveira, C. (2007), EDACLUSTER: Um Algoritmo Evolucionário para Análise de GArupamentos Baseados em Densidade e Grade, *Dissertação (Mestrado em Engenharia Elétrica)*, Universidade Federal do Pará.

Rakesh, A., Johanners, G., Dimitrios, G. & Prabhakar, R. (1999). Automatic subspace clustering of high dimensional data for data mining applications. *In: Proc. of the ACM SIGMOD*, p.94-105.

Spath, H. (1980), Cluster Analysis Algorithms – For Data Reduction and Classification of Objects, *Ellis Horwood Limited.*

Macqueen, J. B. (1967). Some Methods for Classification and Analysis of MultiVariate Observations. *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press. P. 281-297, V. 1.

Mitchell, M. (1999), An Introduction to Genetic Algorithms, *MIT Press*.

Teixeira, O. (2005), Proposta de um novo Algoritmo Genético baseado na Teoria dos Jogos. *Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal do Pará.*

Teknomo, K., Similarity Measurement, *Disponível em: http:\\people.revoledu.com\kardi\tutorial\Similarity*, Acesso em 29 mar. 2010.

University of Waikato, Weka 3 – Machine Learning Software in Java, *Disponível no site da University of Waikato*. URL: http://www.cs.waikato.ac.nz/ml/weka, Acesso em 29 mar 2010.

Whitley, D (1994), A Genetic Algorithm Tutorial, *Computer Science Department*, Colorado State University.