Determinação do Número de Agrupamentos em Conjuntos de Dados Multidimensionais Utilizando Algoritmos Genéticos

Sandro Carvalho Izidoro

Universidade Vale do Rio Verde de Três Corações – UNINCOR – MG
sandroizidoro@gmail.com

Resumo: A análise de agrupamentos tem sido utilizada com sucesso nas mais diversas áreas de pesquisa com o objetivo de agrupar dados semelhantes segundo suas características. Uma técnica eficiente na análise de agrupamentos é a utilização da função de densidade de probabilidade que apresenta o número de agrupamentos graficamente. Os algoritmos genéticos foram utilizados com sucesso para informar o número de agrupamentos em um conjunto de dados unidimensional. Os métodos existentes para a análise de agrupamentos em dados multidimensionais necessitam de um número aproximado de agrupamentos para localizá-los. O desempenho destes métodos depende diretamente deste número de agrupamentos. O propósito deste trabalho é utilizar os algoritmos genéticos para predizer o número de agrupamentos em dados multidimensionais.

Palavras-chave: Inteligência Artificial, Análise de Agrupamentos, Algoritmos Genéticos, Função de Densidade de Probabilidade.

Determination of the number of clusters in multidimensional data sets using Genetic Algorithms

Abstract: The cluster analysis has been used successfully in several research areas with the objective of grouping similar data according to their features. An efficient technique in cluster analysis is the use of the probability density function that presents the number of clusters graphically. The genetic algorithms were used with success to inform the number of clusters in a group of unidimensional data. The existent methods for the cluster analysis in multidimensional data need an approximate number of clusters to locate them. The acting of these methods depends directly on this number of clusters. The purpose of this work is to use the genetic algorithms to predict the number of clusters in multidimensional data.

Keywords: Artificial Intelligence, Cluster Analysis, Genetic Algorithms, Probability Density Function.

(Received August 10, 2005 / Accepted September 15, 2005)

1. Introdução

A análise de agrupamentos ("*clusters*") vem sendo utilizada com sucesso em várias áreas de pesquisa tais como na arqueologia e na biologia, e seu objetivo está em agrupar dados semelhantes segundo suas características, gerando classes. Este tipo de análise tem se demonstrado muito útil no reconhecimento de caracteres, símbolos, figuras, imagens biomédicas, eletrocardiogramas, ondas sísmicas e ondas sonoras.

Existem vários métodos (*hierárquicos* e *não hierárquicos*) para efetuar o agrupamento dos dados, mas para todos os casos existe a necessidade de saber aproximadamente o número de agrupamentos existentes no conjunto de dados analisado. De posse deste número aproximado é que os métodos vão começar a agrupar os dados semelhantes [1].

Uma técnica muito eficiente na análise de agrupamentos é a utilização da função de densidade de probabilidade. Após a estimação da densidade dos dados é possível apresentá-los graficamente. Os picos (*peaks*) desta função representam o número de agrupamentos [4].

Os algoritmos genéticos são algoritmos de busca baseados em mecanismos da seleção natural e genética. Sem as limitações encontradas nos métodos tradicionais, os algoritmos genéticos se mostram muito eficientes para busca de soluções ótimas, ou aproximadamente ótimas, em uma grande variedade de problemas [3].

Os algoritmos genéticos foram utilizados com sucesso para predizer o número de agrupamentos em um conjunto de dados unidimensional [4]. Nenhuma informação foi encontrada na literatura para uma extrapolação para casos bidimensionais e multidimensionais.

O objetivo deste trabalho é o de avaliar a eficiência dos algoritmos genéticos na busca de soluções ótimas para determinar o número de agrupamentos em dados bidimensionais e multidimensionais.

2. Análise de Agrupamentos e Estimação de Densidade

2.1 Introdução

O objetivo principal da análise de agrupamentos está em dividir um determinado conjunto de dados em um número de agrupamentos (*"clusters"*) ou classes. No entanto, não existe nenhuma informação prévia sobre o conjunto de dados. Os dados, sem ajuda, definirão quantos agrupamentos existem e a que regras estarão submetidos nestes agrupamentos.

Várias técnicas de análise de agrupamentos são baseadas em achar semelhanças entre padrões dentro dos dados. Uma técnica muito eficiente é a utilização da função de densidade de probabilidade, onde é possível estimar a densidade dos dados e apresentá-los graficamente. Os picos (*peaks*) desta função representam os agrupamentos [4].

2.2 Estimação de Densidade

Considerando um conjunto aleatório X em uma função de densidade de probabilidade *f*, a função irá fornecer uma descrição da distribuição desse conjunto permitindo encontrar probabilidades associadas com X a partir da equação:

$$P(a < X < b) = \int_a^b f(x)dx \quad \text{para todo } a < b$$

O objetivo da estimação de densidade é construir uma estimativa da função de densidade dos dados em questão uma vez que, freqüentemente, essa função é desconhecida.

Existem dois tipos de estimação de densidade: o paramétrico e o não paramétrico. O primeiro tipo considera que os dados são retirados de um conjunto conhecido, por exemplo: uma distribuição normal com média μ e variância σ^2. Portanto, a estimação de densidade pode ser feita encontrando-se a estimativa de μ e σ^2. O segundo tipo, que será abordado nesse trabalho, considera que os dados são obtidos de um conjunto que não se conhece [6].

O histograma (Figura 1) é o estimador de densidade mais simples e mais usado. A distribuição de densidade de probabilidade é construída através de barras com largura *h* distribuídas ao longo do intervalo onde os dados estão.

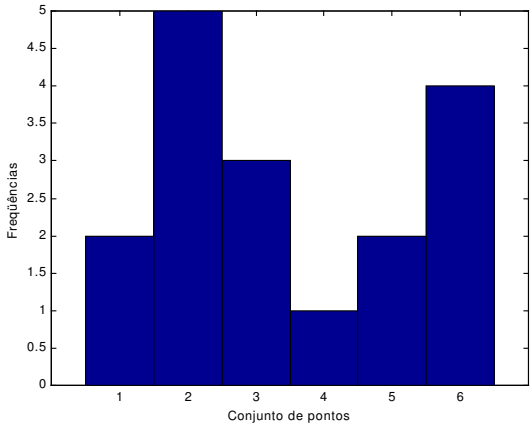

Figura 1 – Exemplo de um histograma

2.3 Utilizando um Estimador Kernel

Um estimador kernel pode ser visto como uma melhoria do histograma e em vez de usar uma função constante, uma função kernel *K* é usada para gerar um *novo histograma*. O estimador kernel é definido por:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - Xi}{h}\right)$$

em que,

X_i – i-ésimo ponto do conjunto de dados do experimento;

x – ponto onde será calculada a função de densidade de probabilidade;

K – uma função escolhida arbitrariamente;

h – coeficiente de suavidade (equivalente a largura dos retângulos no histograma);

n – número de resultados do experimento.

A função kernel *K* pode ser qualquer função de densidade de probabilidade (Gaussiana, Triangular, Retangular, Epanechnikov etc.) desde que satisfaça a seguinte condição:

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

Neste trabalho utiliza-se uma Gaussiana como função kernel *K* por ser mais suave e apresentar os dados de forma mais realista, pois a maioria dos processos analisados apresenta este tipo de distribuição. A Gaussiana como função kernel é definida pela seguinte equação:

$$K(x) = \frac{1}{\sqrt{2*\pi}} e^{\frac{x^2}{2}}$$

Quando se utiliza uma função Gaussiana como função kernel está se colocando uma pequena Gaussiana centrada em cada um dos pontos do conjunto de dados analisado. Posteriormente, soma-se todas as Gaussianas a fim de chegar na função de densidade de probabilidade de todos os pontos (Figura 2).

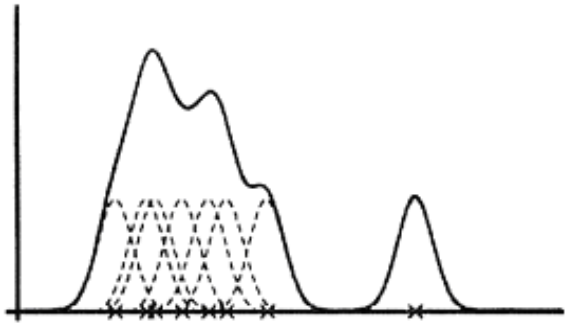

Figura 2 – Função de densidade de probabilidade utilizando estimador kernel. Fonte: Baseado em [4].

2.4 Um Estimador Kernel Multidimensional

A definição de um estimador kernel como um somatório de *picos* centrados em cada um dos pontos do conjunto de dados analisado é facilmente generalizado para o caso bivariado (Figura 3), e conseqüentemente, para o caso multivariado. Para tal, é adotada a notação **x** (*negrito*) para um conjunto de dados multivariados de *d* dimensões. O estimador kernel multivariado é definido por:

$$f(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left\{\frac{1}{h}(\mathbf{x} - \mathbf{Xi})\right\}$$

em que,

Xi – i-ésimo ponto de um conjunto de dados multivariado;

x – ponto onde será calculada a função de densidade de probabilidade;

K – uma função escolhida arbitrariamente;

h – coeficiente de suavidade (equivalente a largura dos retângulos no histograma);

n – número de resultados do experimento;

d – número de dimensões.

A função kernel *K* (**x**) é agora uma função definida por **x** de *d* dimensões satisfazendo a seguinte condição:

$$\int_{R^d} K(\mathbf{x})d\mathbf{x} = 1$$

A Gaussiana como função kernel para o caso multivariado é definida pela seguinte equação:

$$K(\mathbf{x}) = (2\pi)^{-d/2} \exp(-\frac{1}{2}\mathbf{x}^{\mathbf{T}}\mathbf{x})$$

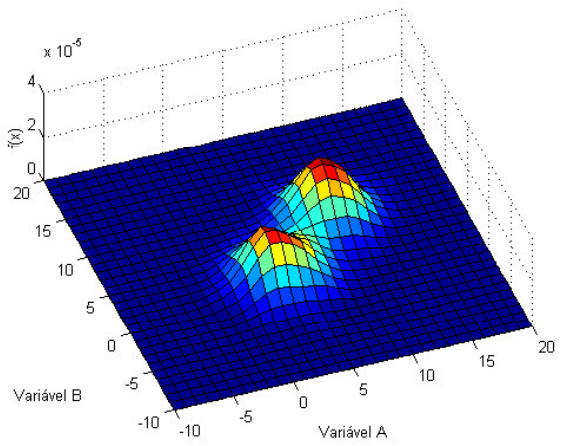

Figura 3 – Exemplo de uma função de densidade de probabilidade utilizando um estimador kernel bivariado.

3. Algoritmos Genéticos

3.1 Definição de Algoritmos Genéticos

Algoritmos Genéticos (AGs) são algoritmos de busca baseados em mecanismos da seleção natural e genética e foram desenvolvidos por John Holland e sua equipe na Universidade de Michigan [2].

Os AGs são muito simples do ponto de vista computacional entretanto são métodos de busca

extremamente eficientes. Partindo de uma *população de candidatos*, os AGs realizam uma busca paralela em diferentes áreas do espaço de soluções. Muito eficientes para busca de soluções ótimas, ou aproximadamente ótimas, em uma grande variedade de problemas, os AGs não impõem muitas das limitações encontradas nos métodos de busca tradicionais [3].

Os AGs são métodos de busca guiados pela função de aptidão. Aleatórios, não executam buscas sem rumo, pois através de processos iterativos (*gerações*) eles exploram informações históricas de cada *geração* para encontrar novos pontos de busca onde são esperados melhores desempenhos [7].

3.2 Conceitos Básicos

Pode-se explicar o funcionamento de um algoritmo genético clássico expondo naturalmente alguns conceitos básicos. O primeiro passo é gerar uma população inicial onde seus indivíduos representam possíveis soluções para um determinado problema. Esta população inicial pode ser gerada a partir de valores aleatórios ou a partir de valores predefinidos (*sementes*). Cada indivíduo é avaliado de acordo com o problema em questão onde os mais aptos são mantidos e os demais são eliminados. Por meio de operadores genéticos (*cruzamento* e *mutação*) os indivíduos restantes geram descendentes (*reprodução*) os quais tem uma grande possibilidade de serem mais aptos do que seus genitores. A *reprodução* é repetida até que uma condição de parada seja satisfeita. Esta condição pode estar relacionada com uma solução satisfatória, o número de gerações ou até mesmo o tempo de processamento.

Em um algoritmo genético clássico um indivíduo é representado por uma *string* binária (***0,1***) onde cada elemento é chamado de *gene* (Figura 4). Cada elemento da *string* pode indicar a presença (***1***) ou ausência (***0***) de uma determinada característica que na genética é referenciada como genótipo. Os elementos combinados formam as características reais do indivíduo ou o seu fenótipo.

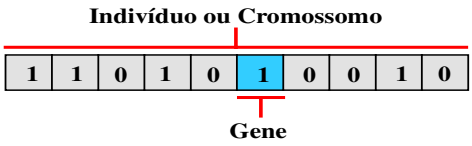

Figura 4 – Representação de um cromossomo de genes binários

3.3 Operadores e Parâmetros Genéticos

A função dos operadores genéticos é, por meio de um processo iterativo, transformar a população inicial em uma população que represente um resultado satisfatório. Um algoritmo genético clássico é composto de três operações [2]:

1 . Reprodução ou Seleção;
2 . Cruzamento;
3 . Mutação.

A idéia básica da reprodução é selecionar os melhores indivíduos da população corrente através de uma função de aptidão. Os indivíduos com um alto valor de aptidão terão uma alta probabilidade de contribuir com um ou mais descendentes na próxima geração.

A operação de reprodução pode ser implementada de várias formas, porém, o método mais utilizado é o método da roleta. Neste método cada indivíduo da população corrente tem sua representação na roleta de acordo com o seu valor de aptidão. Indivíduos com valores de aptidão altos terão um segmento maior dentro da roleta e os indivíduos com valores menores terão segmentos menores (Tabela 1 e Figura 5). Posteriormente a roleta é girada *n* vezes e de acordo com o tamanho da população os indivíduos sorteados é que irão fazer parte da próxima geração.

Tabela 1
Exemplo de uma população com respectivos valores de aptidão

Nº	Indivíduos	Aptidão	% do Total
1	10011	361	21
2	10101	441	26
3	11110	900	52
4	00011	9	1
Total		**1711**	**100**

O operador de cruzamento é utilizado após a reprodução. Nessa fase acontece a troca de segmentos entre casais de indivíduos dando origem a novos indivíduos. Com essa troca o que se tenta fazer é propagar as características dos indivíduos mais aptos da população corrente para futuras gerações.

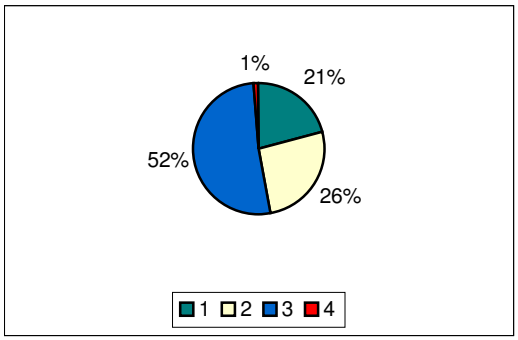

Figura 5 – Roleta de seleção de acordo com os valores de aptidão da Tabela 1

Os indivíduos selecionados pela roleta serão transferidos para uma *piscina de acasalamento* (*"mating pool"*) onde o cruzamento é realizado em dois passos. O primeiro passo consiste em definir os casais de indivíduos de forma aleatória. Em seguida um ponto de quebra do indivíduo é escolhido de forma aleatória ao longo da *string* que o representa. A partir deste ponto é realizada a troca de genes entre o par de indivíduos (Figura 6). Também existem outras formas de efetuar o cruzamento, como a escolha de múltiplos pontos ao longo da string.

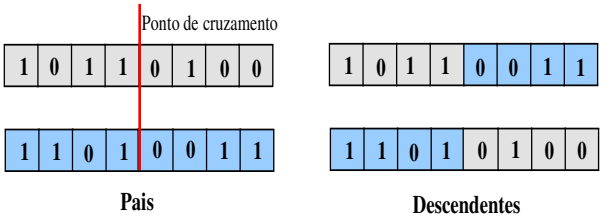

Figura 6 – Operação de cruzamento

Após o cruzamento a operação de mutação é aplicada para cada gene de todos os novos indivíduos de forma aleatória. A operação consiste simplesmente em alterar o valor do gene (1 para 0 e vice versa) (Figura 7). Utilizada para dar uma nova informação para a população, a mutação previne a saturação da população com indivíduos semelhantes.

O operador de mutação garante que a probabilidade de se chegar a qualquer ponto do espaço de busca nunca será zero, além de contornar o problema de mínimos locais.

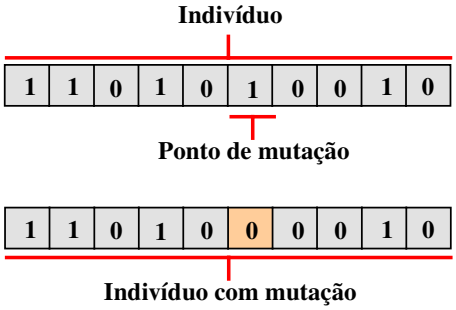

Figura 7 – Operação de mutação

O funcionamento de um algoritmo genético e seus operadores depende de alguns parâmetros que podem ser ajustados conforme as necessidades de cada problema. Os parâmetros utilizados são:

- Taxa de Cruzamento: valor que determina a probabilidade de cruzamento dentro de uma população;
- Taxa de Mutação: determina a probabilidade que uma mutação ocorrerá;
- Tamanho da População.

3.4 Codificação dos Indivíduos

Outro ponto importante é a codificação dos indivíduos. Quando se deseja que o próprio indivíduo (string binária) represente um valor numérico, a técnica mais utilizada é através de uma representação discreta dos dados dentro de um intervalo [x_{min}, x_{max}] em uma quantidade de pontos 2^t, tal que à distância entre pontos consecutivos seja menor que um valor de tolerância especificado, ou seja [5]:

$$\frac{x_{\max} - x_{\min}}{2^t - 1} < TOL$$

Portanto, cada ponto do espaço de busca será representado por um número binário de tamanho *t*, começando por 0...0 que representa x_{min} e terminando por 1...1 que representa x_{max} .

O ponto principal da representação está em calcular o tamanho (*t*) dos indivíduos. Com base no valor de tolerância este tamanho pode ser calculado a partir da seguinte equação [5]:

$$t = \log_2\left(1 + \frac{x_{\max} - x_{\min}}{TOL}\right)$$

Por exemplo, em um intervalo x ∈ [0,1] e uma precisão de duas casas decimais (TOL = 5 x 10^{-3}), então o tamanho do indivíduo seria:

$$t = \log_2\left(1 + \frac{1-0}{0.005}\right) = 8$$

Desta forma, pode-se utilizar indivíduos com o tamanho de 8 bits para representar o intervalo x ∈ [0,1] com precisão menor igual a 0.005.

4. Análise de Agrupamentos Utilizando Algoritmos Genéticos

4.1 Introdução

A técnica de análise de agrupamentos utilizando AGs é muito simples e eficiente. A idéia principal consiste em utilizar os AGs para encontrar os máximos da função de densidade de um conjunto de dados [4].

O algoritmo genético para esta implementação utiliza o estimador kernel como função de aptidão. O objetivo é achar todos os máximos locais obtidos pela função de aptidão uma vez que o agrupamento dos dados não está apenas no máximo global da função de densidade. Uma das características dos AGs é que eles podem achar os máximos locais de um conjunto de dados a partir de uma população pequena com um número pequeno de gerações [5].

A execução do algoritmo genético em apenas uma vez não garante que todos os máximos locais serão encontrados. Executa-se o algoritmo genético *N* vezes, armazenando a população final após *M* gerações. Em seguida, calcula-se a função de densidade da população de soluções e os picos apresentados representarão os agrupamentos.

4.2 Análise de Agrupamento Utilizando AGs para Dados Unidimensionais

Para avaliar e exemplificar a utilização dos algoritmos genéticos na análise de agrupamentos, foi primeiramente desenvolvida uma aplicação para dados unidimensionais. Foram gerados três subconjuntos de dados com uma distribuição normal (Gaussiana) baseados nos parâmetros apresentados na Tabela 2. Os subconjuntos foram gerados utilizando o software MATLAB. Os três subconjuntos formam o conjunto de dados que será analisado. O que se espera é encontrar três agrupamentos centrados nos pontos 2, 7 e 11 conforme a média de cada subconjunto apresentado na Tabela 2.

A Figura 8 apresenta a função de densidade dos dados estimada utilizando apenas o estimador kernel.

Tabela 2
Parâmetros do conjunto de dados unidimensional

Subconjuntos	Média	Variância	Nº de Pontos
1	2	2	600
2	7	2	600
3	11	2	600
Total	**-**	**-**	**1800**

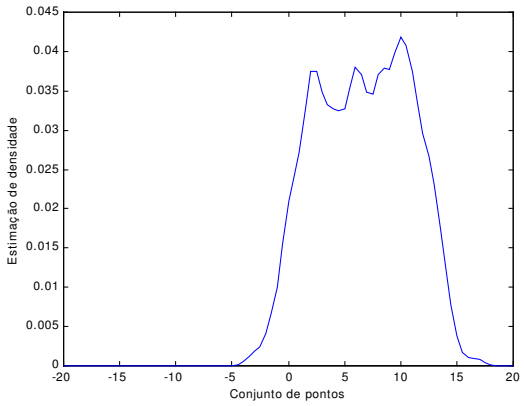

Figura 8 – Função de estimação de densidade utilizando o estimador kernel

O passo seguinte é realizar a estimação de densidade utilizando os AGs. Os passos desta técnica são:

- Definir o estimador kernel como função de aptidão;
- Definir uma pequena população inicial;
- Definir um valor pequeno para o número máximo de gerações;
- Executar o algoritmo genético *N* vezes e salvar a população final a cada vez;
- Utilizar o estimador kernel para estimar a função de densidade da população final obtida depois da execução do algoritmo genético *N* vezes;

- O número de picos será o número de agrupamentos e as variáveis de cada pico serão o centro dos agrupamentos.

Os indivíduos foram codificados conforme o item 3.4, com tamanho de 21 bits, e uma precisão de 0,00001.

O programa (implementado na linguagem C) é executado cem vezes e a cada dez gerações ele armazena a população final. A função de densidade deste conjunto é estimada e o resultado é apresentado na Figura 9.

Analisando os resultados percebe-se claramente o poder dos AGs. Em uma comparação visual observa-se que a Figura 9 é mais clara e precisa que a Figura 8 que foi gerada apenas com o estimador kernel.

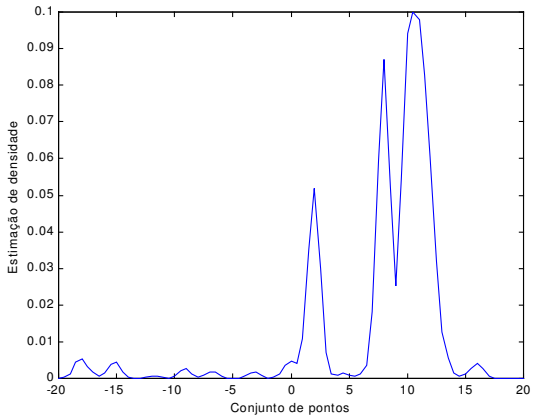

Figura 9 – Função de estimação de densidade utilizando a população final do algoritmo genético

5. Análise de Agrupamentos Utilizando Algoritmos Genéticos para Dados Multidimensionais

5.1 Introdução

O objetivo deste trabalho é utilizar os algoritmos genéticos para determinar o número de agrupamentos em dados bidimensionais e multidimensionais. Para isto, algumas modificações deverão ser feitas no algoritmo genético que foi utilizado para dados unidimensionais. A nova implementação utilizará o estimador kernel multivariado como função de aptidão. Os parâmetros utilizados no algoritmo genético também deverão ser modificados (taxas de cruzamento, mutação, número de gerações, etc.). Outra alteração importante diz respeito ao tratamento que o algoritmo dará aos dados analisados. Para dados bidimensionais, por exemplo, a nova aplicação terá um conjunto de N dados da seguinte forma:

(x1, y1), (x2, y2), (x3, y3), ... (xN, yN)

Cada par ordenado (x*i*, y*i*) deverá ter um valor para a função de densidade de probabilidade.

O algoritmo genético não pode tratar os dados separadamente. Após a aplicação dos operadores genéticos (*cruzamento* e *mutação*) em um conjunto de dados bidimensional um valor de x pode *evoluir* e o mesmo pode não acontecer com um valor de y. A idéia é concatenar os indivíduos de x e y, gerando um novo e único indivíduo (Figura 10). Assim os operadores genéticos serão aplicados no par (x*i*, y*i*) como um todo e a evolução será do par e não apenas de uma das variáveis. A técnica é a mesma para os casos multidimensionais, ou seja, todas as variáveis (*indivíduos*) devem ser concatenadas para gerar um único indivíduo.

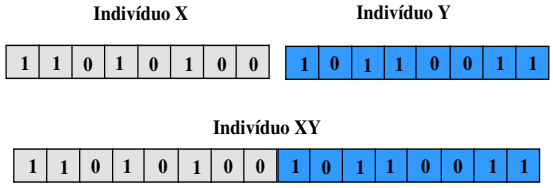

Figura 10 – Concatenação dos indivíduos x e y

5.2 Caso Bivariado

O algoritmo genético para esta implementação utiliza o estimador kernel multivariado para duas variáveis como função de aptidão.

No caso univariado o algoritmo genético foi configurado com um pequeno número de gerações. Para o caso bivariado este número deve ser aumentado. Para trabalhar com duas variáveis o algoritmo genético precisa de mais gerações para conseguir evoluir sua população. As taxas de cruzamento e mutação também foram alteradas para permitir que o algoritmo genético possa atingir todo o espaço de soluções. Os passos desta técnica foram definidos da seguinte maneira:

- Definir o estimador kernel multivariado como função de aptidão;

- Definir duas populações iniciais de tamanho pequeno (*x* e *y*);
- Definir um valor pequeno para o número máximo de gerações;
- Concatenar os indivíduos de *x* e *y* gerando um único indivíduo;
- Executar o algoritmo genético *N* vezes;
- Separar o indivíduo da população final obtendo novamente os indivíduos de *x* e *y*;
- Salvar a populações finais (*x* e *y*) a cada vez;
- Utilizar o estimador kernel multivariado para estimar a função de densidade das populações finais obtidas (*x* e *y*) depois da execução do algoritmo genético *N* vezes;
- O número de picos será o número de agrupamentos e as variáveis de cada pico serão o centro dos agrupamentos.

A codificação dos indivíduos foi à mesma utilizada para dados unidimensionais.

Para testar o algoritmo genético para o caso bivariado foi utilizado um conjunto de dados gerados conforme a Tabela 3.

No conjunto de dados tem-se para as variáveis A e B médias 1, 5 e 10. Os dados foram colocados aos pares, e assim, espera-se encontrar três agrupamentos nas posições 1 e 1, 5 e 5, e 10 e 10.

Tabela 3
Parâmetros do conjunto de dados bidimensional

Variáveis	Média	Variância	N° de Pontos
A	1	1	300
	5	1	300
	10	1	300
B	1	1	300
	5	1	300
	10	1	300

O resultado apresentado na Figura 11 foi obtido utilizando apenas o estimador kernel multivariado.

O resultado apresentado na Figura 12 foi obtido utilizando o estimador kernel multivariado e o algoritmo genético. Pode-se observar que os agrupamentos apresentados na Figura 11 também foram encontrados na Figura 12.

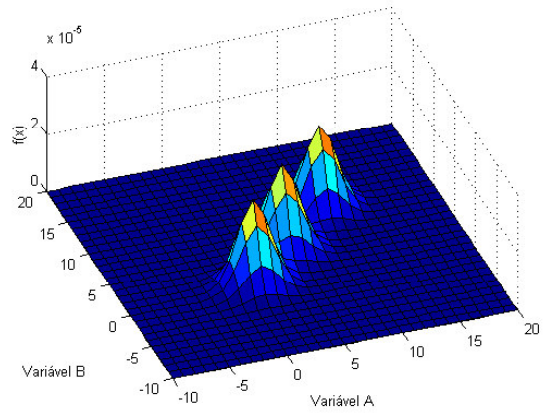

Figura 11 – Estimador kernel multivariado baseado nos dados da Tabela 3

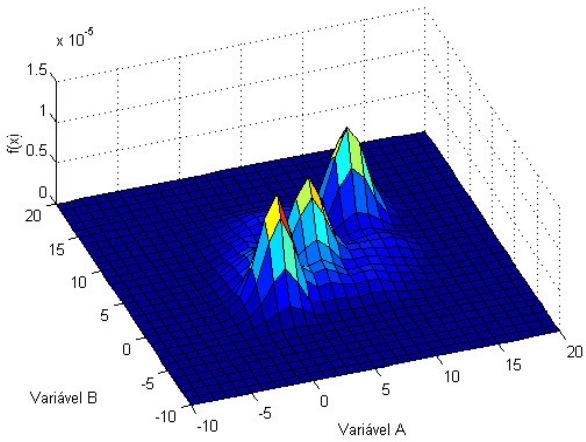

Figura 12 – Estimador kernel multivariado com algoritmo genético baseado nos dados da Tabela 3

5.4 Caso Multivariado

A técnica para o caso multivariado será a mesma utilizada para o caso bivariado com pequenas alterações.

Para os testes o algoritmo genético foi implementado utilizando o estimador kernel multivariado para três variáveis. Por causa deste aumento de variáveis o número de gerações deverá ser aumentado em relação ao caso bivariado.

A outra alteração está relacionada à geração dos gráficos. Não é possível gerar um gráfico com mais de duas variáveis. Neste caso será utilizado um gráfico

unidimensional onde *os picos* apresentados indicarão *o número de agrupamentos*.

Os passos desta técnica foram definidos da seguinte maneira:

- Definir o estimador kernel multivariado como função de aptidão;
- Definir três populações iniciais de tamanho pequeno (*x, y* e *z*);
- Definir um valor pequeno para o número máximo de gerações;
- Concatenar os indivíduos de *x, y* e *z* gerando um único indivíduo;
- Executar o algoritmo genético *N* vezes;
- Salvar a população final (*xyz*) a cada vez;
- Utilizar o estimador kernel univariado para estimar a função de densidade das população final obtida (*xyz*) depois da execução do algoritmo genético *N* vezes;
- O número de picos será o número de agrupamentos.

Para testar o algoritmo genético para o caso multivariado foi utilizado um conjunto de dados (Figura 13) gerado conforme a Tabela 4. Foram geradas quatro médias para cada variável. Os dados foram emparelhados, e assim, espera-se encontrar quatro agrupamentos.

Tabela 4
Parâmetros do conjunto de dados multidimensionais

Variáveis	Média	Variância	Nº de Pontos
A	1	1	50
	9	1	50
	5	1	50
	13	1	50
B	5	1	50
	5	1	50
	1	1	50
	8	1	50
C	3	1	50
	7	1	50
	1	1	50
	1	1	50

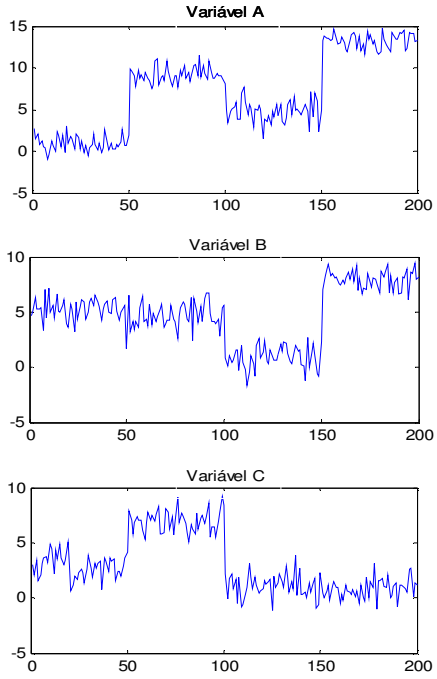

Figura 13 – Conjunto de dados multidimensional conforme Tabela 4

Na Figura 14 é apresentado o resultado do estimador kernel multivariado com algoritmo genético. Pode-se perceber claramente quatro picos correspondentes aos quatro agrupamentos esperados.

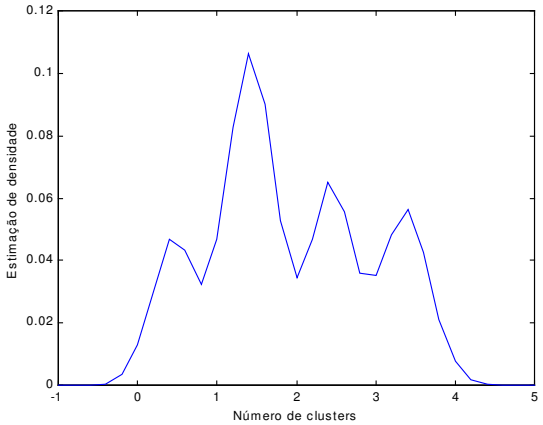

Figura 14 – Estimador kernel multivariado com algoritmo genético baseado nos dados da Tabela 4

6. Conclusão

A aplicação dos algoritmos genéticos na análise de agrupamentos se mostrou eficiente.

Em dados unidimensionais e bidimensionais os algoritmos genéticos encontraram o mesmo número de agrupamentos que o estimador kernel.

No caso multidimensional, os algoritmos genéticos encontraram o número esperado de agrupamentos, apontando uma alternativa para o problema da falta de informação sobre o número de agrupamentos existentes no conjunto de dados a ser analisado.

Os métodos para a análise de agrupamentos em dados multidimensionais utilizam um número aproximado de agrupamentos. Baseado neste número os métodos localizam os agrupamentos no conjunto de dados analisado. O número de agrupamentos agora pode ser conseguido utilizando os algoritmos genéticos e dessa forma conseguir um melhor resultado na localização e agrupamento dos dados.

7. Referências Bibliográficas

[1] Bow, S. *Pattern Recognition:* Application to Large Data-Set Problems. Marcel Dekker, Inc., 1984. 323p.

[2] Goldberg, D. E. *Genetic algorithms in search, optimization, and machine learning*. Reading: Addison Wesley, 1989. 412p.

[3] Mendes Filho, E. F. *Algoritmos genéticos* [online]. 1998. Disponível na Internet via World Wide Web: <http://www.icms.sc.usp.br/~prico/index.html>.

[4] Pinto, J. O. P. *Cluster Analysis using GA*, University of Tenessee, 1998.

[5] Serrada, A. P. *Una introducción a la computación evolutiva* [online]. 1996. Disponível na Internet via World Wide Web: <http://www.geocities.com/ igoryepes/>.

[6] Silverman, B. W. *Density estimation for statistics and data analysis* (Monographs on statistics and applied probability). London: Chapman and Hall, 1990. 175p.

[7] Yepes, I. *Uma incursão aos algoritmos genéticos* [online]. Disponível na Internet via World Wide Web: <http://www.geocities.com/igoryepes/ >.