UNIVERSIDADE ESTADUAL DO CEARÁ
CENTRO DE CIÊNCIA E TECNOLOGIA
MESTRADO ACADEMICO EM CIÊNCIA DA COMPUTAÇÃO

FÁBIO DOS SANTOS FERREIRA

ESTIMATIVA DA INFLUÊNCIA DE PONTOS ATRATORES EMPREGANDO-SE ALGORITMOS GENÉTICOS E FUNÇÕES DENSIDADE DE INFLUÊNCIA

FORTALEZA - CEARÁ
2011

FÁBIO DOS SANTOS FERREIRA

ESTIMATIVA DA INFLUÊNCIA DE PONTOS ATRATORES EMPREGANDO-SE ALGORITMOS GENÉTICOS E FUNÇÕES DENSIDADE DE INFLUÊNCIA

Dissertação apresentada ao Curso de Mestrado em Ciência da Computação do Centro de Ciência e Tecnologia da Universidade Estadual do Ceará, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação

Orientadores: Prof. Dr. Jackson Sávio de Vasconcelos Silva e Prof. Dr. Gustavo Augusto Lima de Campos

FORTALEZA - CE
2011

FÁBIO DOS SANTOS FERREIRA

ESTIMATIVA DA INFLUÊNCIA DE PONTOS ATRATORES EMPREGANDO-SE ALGORITMOS GENÉTICOS E FUNÇÕES DENSIDADE DE INFLUÊNCIA

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Estadual do Ceará, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Aprovado em: 09/08/2011

BANCA EXAMINADORA

_____

Prof. Dr. Jackson Sávio De Vasconcelos Silva (Orientador)
Universidade Estadual do Ceará - UECE

_____

Prof. Dr. Gustavo Augusto Lima de Campos (Co-Orientador)
Universidade Estadual do Ceará - UECE

_____

Prof. Dr. Antônio Clécio Fontelles Thomaz
Universidade Estadual do Ceará - UECE

_____

Prof. Dr. Marcelino Pereira dos Santos Silva
Universidade Federal Rural do Semi-Árido - UFERSA

AGRADECIMENTOS

Agradeço aos meus pais, **Benedito Aldolfo** e **Rosinete dos Santos**, pelos ensinamentos que me deram, por acreditarem e me darem força em vários momentos da minha vida, e agora principalmente no período em que permaneci distante de casa.

As minhas irmãs, **Kelly Cristina** e **Paula Cristina** por estarem na minha, vida me ajudando em vários momentos.

A minha namorada e companheira **Monik Kely**, por estar ao meu lado e me dar o apoio que preciso, sem ela não teria chegado tão longe.

Aos meus tios e avós, que sempre me apoiaram em vários momentos.

Agradeço a **Maria José**, e sua família. Por me acolher como parte da família e ser uma verdadeira mãe para mim durante o tempo em que permaneci em Fortaleza.

Aos meus amigos de Belém **Glauber Duarte**, **Jesus Nazareno**, **André Furtado**, **David Pinheiro**, **Helton José** por estarem seguinte em frente em busca dos sonhos de uma equipe.

Aos meus amigos de Fortaleza **Antoniel Rego**, **Alyson de Oliveira**, **Gustavo Sikora**, **Walisson Pereira**, **Marco Antonio**, **Francisco Vando** pelos momentos de descontração e apoio.

Aos meus amigos e professores na graduação **Otávio Noura**, **Alessandra Natasha e Marcos Venicius** por me apoiarem na decisão de cursar o mestrado em computação.

Aos meus amigos e orientadores no mestrado **Prof. Dr. Jackson Sávio** e **Prof. Dr. Gustavo Augusto** por acreditarem em mim e me indicar o caminho correto durante o curso, sem eles eu não conseguiria.

A todos os professores do Mestrado Acadêmico da Universidade Estadual do Ceará pela ajuda e pelos ensinamentos durante o curso.

Agradeço também ao **Governo Federal** por criar o programa **PROUNI** (Programa Universidade para Todos), sendo este o motivo de estar graduando pelo **CESUPA**.

Pelo apoio do **CNPQ** por me conceder a bolsa que me manteve durante os anos em que estive em fortaleza.

Ser racional é levar em consideração as conseqüências de suas ações e saber que o resultado destes pode depender da conseqüência das ações de outros agentes.

Anatol Rapoport

RESUMO

Este trabalho apresenta uma técnica de análise de pontos de influência em bases de dados. Esta técnica utiliza um algoritmo fundamentado em um primeiro algoritmo genético (AG) e noções de densidade e grade para agrupar os dados das bases, e em um segundo algoritmo genético e funções de influência e densidade de influência para o cálculo de pontos atratores nas regiões da grade que contém os agrupamentos. Os experimentos realizados confirmaram a qualidade dos agrupamentos gerados pelo primeiro AG, bem como a eficiência do segundo no cálculo dos pontos atratores, e a utilidade da descrição de uma função densidade de influência nos pontos atratores e a descrição de como esta influência é atenuada à medida em que os pontos estão mais distantes de um atrator em consideração.

Palavras-Chaves: Algoritmo Genético, Funções de Influência, Densidade, Grade, Agrupamento, Atratores de Densidade

ABSTRACT

This assignment presents an analysis of points of influence in databases. This technique uses an algorithm based on a first genetic algorithm (GA) and notions of density and grid to group the data bases, and a second genetic algorithm and influence functions and density of influence for the calculation of attractors points in the regions of grid containing clusters. The experiments confirmed the quality of clusters generated by the first AG as well as efficiency of the second in the calculation of point attractors, and the utility of description of a density function of influence in the points attractors and description of how this influence is mitigated to the extent that the points are farther of an attractor into account.

Keywords: Generic Algorithm, Influence Function, Density, Grid, Cluster, Density Attractors.

LISTA DE SIGLAS

AD	Agrupamento de Dados
AE	Algoritmos Evolutivos
AG	Algoritmo Genético
AGABDG	Algoritmo Genético para Agrupamento Baseado em Densidade e Grade
CLIQUE	*Clustering In Quest* – Análise dos Componentes Principais
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
DENCLUE	*Density-Based Clustering*
EDA	Algoritmo de Estimativa de Distribuição - *Estimation of Distribution Algorithms.*
IF	*Influence Functions* – Funções de Influencia
KDE	*Kernel* Density Estimation
PCA	*Principal Components Analysis*
STING	*Statistical Information Grig*

SUMÁRIO

1 INTRODUÇÃO

Estimadores de densidade auxiliam na investigação formal das propriedades de um conjunto de dados. O seu estudo pode gerar valores indicativos da assimetria e multimodalidade dos dados (SILVERMAN, 1986). Analisando a densidade estimada é possível encontrar regiões de alta densidade, sendo que no centro desta região é encontrado o ponto que possivelmente a atrai.

***Kernel* Density Estimation** (KDE) é um método de estimação de densidade que utiliza uma função chamada *Kernel function* para calcular a influência de um ponto sobre outro e a influência de um conjunto de pontos sobre um ponto.

A partir da distancia entre dois pontos é possível calcular o valor da influencia de um ponto sobre o outro. Um ponto que está próximo de outro sofrerá maior influência do que de outros pontos que estão mais distantes. Através desta função é possível identificar a suavização da influencia à medida que a distancia do ponto de origem é aumentada. Usualmente a função *Kernel* escolhida é uma função de densidade unimodal que é simétrica sobre zero.

A função de *kernel* pode ser utilizada para calcular a densidade de um ponto, através da soma das influencias de um conjunto de pontos sobre este. Caso esta densidade seja a máxima para a sua vizinhança, então este será chamado de atrator de densidade.

A metaheurística de otimização Algoritmo Genético foi utilizada como ferramenta de cálculo destes pontos atratores.

Como método de auxílio para o cálculo de pontos atratores é utilizado um algoritmo de agrupamento desenvolvido. Este agrupamento foi desenvolvido com o intuito de dividir a base de dados em classes, com o intuito de o cálculo do ponto atrator ser realizado sobre este agrupamento. Com isso, será encontrado o atrator de densidade da classe como um todo.

Segundo OLIVEIRA (2007), o agrupamento de dados pode ser definido 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).

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). Neste trabalho foi utilizado um método desenvolvido que utiliza Algoritmos Genéticos com uma forma híbrida dos métodos de densidade e grade de agrupamento.

O Algoritmo Genético (AG) – *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 AG, "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.

Para o agrupamento foi utilizada um algoritmo desenvolvido com o intuito de facilitar a classificação dos dados. Neste agrupamento o fator mais importante é a densidade dos dados, visto que o fator a ser considerado no cálculo dos atratores e esta organização. Por isso, um dos métodos de agrupamento utilizado é por densidade. Outro fator é a localização espacial destes, visto que o atrator será um ponto qualquer no espaço de *features*, muito provavelmente ele não estará entre os dados considerados. Mais sim será calculado a partir da posição destes. O método de grade é utilizado como forma de dividir um subespaço do espaço de *features*, onde o atrator será encontrado.

1.1 OBJETIVOS

A pesquisa realizada neste trabalho busca o desenvolvimento de uma metodologia utilizando Algoritmo Genético (AG) para calcular os pontos atratores de densidade em base de dados. Utilizando um algoritmo de agrupamento baseado em densidade e grade, que também utiliza AG, para fins de busca por estes pontos atratores em classes de dados. Depois de calculados estes pontos atratores, será realizado o calculo de sua influência com relação a classe a qual pertence, e assim estudar a atenuação de sua densidade e influência.

1.1.1 Objetivo geral

De forma geral, a pesquisa tem por objetivo um sistema que possa encontrar pontos atratores na base de dados com o uso da Metaheurística Algoritmo Genético.

1.1.2 Objetivo específico

Especificamente deseja-se, com a abordagem proposta, introduzir uma forma de calcular os pontos atratores de densidade da base de dados. Assim como estudar a sua região de influencia, a partir das funções de densidade de influencia e função *kernel*, para fins de estudo de sua atenuação de densidade e influência, à medida que se distancie do ponto atrator encontrado.

1.2 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação é organizada da seguinte forma: no capítulo 2 é apresentado uma descrição detalhada sobre análise de agrupamento, os principais métodos que podem ser utilizados, os passos para uma tarefa de agrupamento, os requisitos, aplicabilidade e alguns métodos mais utilizados. No capítulo 3 é apresentada uma descrição sobre o método de Funções de Influencia, que serão utilizados para encontrar os pontos atratores na base de dados e serão utiliados para calcular a sua densidade. No capítulo 4 é feito um detalhamento sobre Algoritmos Genéticos, e as principais formas de utilizá-lo, descrevendo deus passos e indicando formas opcionais de implementações. No capítulo 5 o Algoritmo Genético para Agrupamento Baseado em Densidade e Grade (AGABDG) é descrito, e todos os passos necessários para a sua utilização, assim como a forma como este foi utilizado neste trabalho

são mostrados. No capítulo 6 o método de calculo de pontos dos atratores de densidade é apresentado. Enfim o capítulo 7 apresentará gráficos com os resultados encontrados para os atratores encontrados utilizando o método proposto no capítulo 6. Bases de dados publicas foram utilizadas, a base Iris e a base Glass. E por ultimo, no capítulo 8, as conclusões serão feitas.

2 AGRUPAMENTO DE DADOS

O objetivo do **Agrupamento de Dados** (AD) é a partição de um conjunto de dados em grupos de dados "similares". É o processo de agrupar um conjunto de objetos físicos ou abstratos em classes similares, chamados grupos, de forma que objetos em um grupo possuam alta "similaridade" em comparação a algum outro, mas sejam "dissimilares" a objetos em outro grupo (HAN & KAMBER, 2006). A dissimilaridade é avaliada baseada nos valores dos atributos que descrevem o objeto. Os objetos em um conjunto de dados podem ser clientes de um banco, figuras ou coisas em uma fotografia, pessoas doentes e saudáveis.

A tarefa de agrupamento pode ser chamada também de **Segmentação de Dados** em algumas aplicações, devido aos grupos particionarem um grande conjunto de dados de acordo com a sua similaridade. O agrupamento também pode ser usado na detecção de pontos isolados, onde os dados isolados podem ser mais interessantes em alguns casos (HAN & KAMBER, 2006).

Nesta linha de pesquisa alguns trabalhos podem ser citados.

Em MAULIK & BANDYOPADHYAY (2000), o algoritmo de clusterização *k-means* (Macqueen, 1967) é utilizado com um Algoritmo Genético graças à sua simplicidade. Na fase de avaliação do cromossomo, no cálculo do *fitness* no GA é utilizado o cálculo 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 AG. 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 são submetidos a uma fase

de união, onde são 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.

A seguir serão apresentados alguns métodos que podem ser utilizados em algoritmos de agrupamento. Diferentes métodos podem gerar diferentes formas para os grupos.

2.1 MÉTODOS DE AGRUPAMENTO

Existem diversos algoritmos de agrupamento na literatura (HAN & KAMBER, 2006; MACQUEEN, 1967; WANG et al., 1997; ESTER et al., 1996; HINNEBURG & KEIM, 1998)**,** muitas vezes classificar e agrupar algoritmos não é uma tarefa simples. Alguns autores já utilizaram formas de classificação através do método utilizado para formar os grupos e testar os resultados, como o encontrado em SPATH (1980 apud Green and Tull, 1970). A seguir são listadas algumas técnicas de análise de agrupamento também encontrado em HAN & KAMBER (2006).

2.1.1 Métodos baseados em particionamento

Em uma base de dados de tamanho *n*, um método particionador procura separar a base de dados em *k* partições, onde cada partição representa um grupo e $k \leq n$ (HAN & KAMBER, 2006). Segundo Han & Kamber essa divisão deve seguir aos seguintes critérios:

- Cada grupo deve conter pelo menos um objeto.
- Cada objeto deve pertencer a exatamente um único grupo. A técnica de particionamento baseada em lógica difusa é uma exceção a esta regra, onde um

objeto pode pertencer a um grupo com maior intensidade, ao mesmo tempo em que pertence a outro grupo com menor intensidade.

Esta técnica parte de uma partição inicial em *k* grupos, e a partir de então movimenta o objeto de um grupo a outro por meio de alguma heurística. Heurísticas deste método podem ser de dois tipos, baseadas em Centróide e baseada em Objeto Representativo

2.1.1.1 Técnica Baseada em Centróide

A partir de *n* objetos a serem agrupados em *k* grupos, a partir de um agrupamento inicial, calculam os pontos médios dos grupos e iteram realocando objetos entre os grupos. Baseando-se nas distâncias entre os objetos e esses pontos médios, tenta maximizar as distâncias entre os pontos médios dos grupos e minimizar o valor médio da soma das distancias entre elementos de um mesmo grupo (NOVAES, 2002). Um algoritmo que implementa esta técnica é o ***K-Means*** (MACQUEEN, 1967).

2.1.1.2 Técnica Baseada em Objeto Representativo

A distância é calculada em relação ao elemento do conjunto mais próximo ao ponto médio. Este método de agrupamento funciona bem para encontrar grupos em formatos esféricos em bancos de dados de pequeno e médio porte. Os métodos de **Particionamento** baseados nesta heurística precisam ser modificados para encontrar grupos com formas complexas e agrupar grandes conjuntos de dados. (HAN & KAMBER, 2006).

2.1.2 Métodos baseados em 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; OLIVEIRA, 2007). Um exemplo de algoritmo baseado em **Densidade** é o algoritmo **DBSCAN (***Density-Based Spatial Clustering of Applications with Noise)* (ESTER ET. Al., 1996) e o **DENCLUE (***Density-Based Clustering)* (HINNEBURG & KEIM, 1998).

2.1.3 Métodos baseados em estruturas 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 faz com que o desempenho dos algoritmos baseados nesta heurística seja bom (HAN & KAMBER, 2006). Alguns exemplos típicos da abordagem baseada em grade incluem **STING (***Statistical Information Grig)* (WANG ET. AL., 1997), que explora a informação estatística armazenada nas células da grade e **CLIQUE** (*Clustering In Quest*) (RAKESH ET. AL., 1999), uma abordagem baseada em **Grade** e em **Densidade** para o agrupamento em espaço de dados de alta dimensão.

Mesmo que os métodos sejam conhecidos e bem entendidos, para uma boa tarefa de análise de agrupamentos é necessário aplicar alguns passos. O seguinte tópico apresenta estes passos.

2.2 PASSOS DE UMA TAREFA DE AGRUPAMENTO

Para que uma boa tarefa de agrupamento seja realizada é necessário, primeiramente, um planejamento inicial. Este planejamento deve definir quais parâmetros serão utilizados pelo sistema que irá realizar a análise de agrupamento, também deve ser feita a preparação dos dados que serão estudados.

Esta preparação envolve a **Limpeza de Dados e Pré-processamento**. Nesta preparação se faz necessário realizar tarefas como remoção de ruídos quando necessário, coleta da informação necessária para modelar ou estimar ruído, escolha de estratégias para manipular campos de dados ausentes, formatação de dados de forma a adequá-los à ferramenta de agrupamento. Após esta preparação o planejamento de como o método de agrupamento irá trabalhar deve ser feito de acordo com os seguintes passos básicos definidos em THEODORIDIS & KOUTROUMBAS (2003).

Seleção dos atributos: Atributos devem ser devidamente selecionadas de modo a codificar o máximo de informações possíveis sobre a tarefa de interesse. A fim de minimizar a redundância de informação entre as características e minimizar o processamento realizado.

Como na classificação supervisionada, o pré-processamento de atributos pode ser necessário antes da sua utilização nas etapas subseqüentes.

Medida de Proximidade: Esta é uma medida que quantifica o quanto “similar” ou “diferentes” são dois vetores de atributos. É natural garantir que todos os elementos selecionados contribuam igualmente para o cálculo da medida de proximidade e que não haja elementos que dominem outros.

Critério de Agrupamento: Isto depende da interpretação que o especialista da ao termo "sensível", baseado do tipo de agrupamentos que são esperados formarem da base de dados. Por exemplo, um agrupamento compacto de objetos em um espaço *d*-dimensional pode ser sensível de acordo com algum critério, enquanto que um agrupamento largo pode ser sensível de acordo com outro. O critério de agrupamento pode ser expresso por uma função de custo ou algum outro tipo de regra.

Algoritmo de Agrupamento: Tendo adotado uma medida de proximidade e um critério de agrupamento, esta etapa se refere à escolha de um algoritmo específico que desvenda a estrutura de agrupamento da base de dados.

Validação dos Resultados: Uma vez que os resultados do algoritmo de agrupamento foram obtidos, é necessário verificar a sua exatidão. Isso geralmente é realizado através de testes adequados.

Interpretação dos Resultados: Em muitos casos, o especialista no campo da aplicação deve integrar os resultados do agrupamento com outras evidências experimentais e análise, a fim de tirar as conclusões corretas.

A escolha de características diferentes para as medidas de proximidade, o critério de agrupamento e algoritmos podem conduzir a resultados totalmente diferentes.

Um passo importante, se não o mais importante, não citado anteriormente é quanto à escolha do número de grupos que o algoritmo irá analisar. A seguir é apresentado como esta escolha pode ser feita.

2.3 ESCOLHA DO NÚMERO DE GRUPOS

Na maioria dos algoritmos de agrupamentos o número de grupos *k* que será encontrado deve ser definido previamente. Em alguns casos este parâmetro pode ser utilizado também como mecanismo capaz de reduzir a complexidade da tarefa.

A seguir serão apresentados alguns meios de realizar uma sintonia fina sobre o valor de *k* (ALPAYDM, 2004):

- Em algumas aplicações, tais como quantização de cor, *k* é definido pela aplicação;

- Plotagem dos dados em duas dimensões utilizando **Análise dos Componentes Principais** (**PCA** – *Principal Components Analysis*) (MANLY, 1994) pode ser utilizada para descobrir a estrutura dos dados e o número de grupos nos dados;

- Uma abordagem incremental também pode ajudar: Configurar uma distância máxima permitida é equivalente a definir um erro máximo de reconstrução permitido, por exemplo;

- Em algumas aplicações, a validação dos grupos pode ser feito manualmente, verificando se realmente o código dos grupos encontrados são realmente significativos. Por exemplo, em uma aplicação de mineração de dados, especialistas podem fazer essa verificação. Na quantização de cor, é possível inspecionar a imagem visualmente para verificar a sua qualidade.

Neste trabalho é desenvolvido um algoritmo de agrupamento que utiliza as metodologias de densidade e grade. Este algoritmo é capaz de trabalhar com atributos de dados numéricos, já que é utilizado o espaço de dados onde estes estão inseridos; é capaz de lidar com alta dimensionalidade, apesar de este trabalhar utilizando todas as dimensões dos dados, é possível utilizar algum método de seleção de atributos; capaz de trabalhar com bases de dados tanto grandes como pequenas, pois utiliza uma etapa de configuração, onde são construídas células onde os objetos serão inseridos. É capaz de descobrir grupos de diferentes formatos, já que este algoritmo trabalha com o agrupamento baseado em densidade, utilizando as células dos objetos, para que este seja utilizado no calculo dos pontos atratores.

3 FUNÇÕES DE INFLUÊNCIA

Segundo PELTONEN & KOUSMANEN (2001), a estatística robusta moderna para estudo de erros em base de dados teve início em 1960 com os trabalhos de TUKEY (1960) & ANSCOMBE (1960). A primeira teoria abordando este assunto foi introduzida por HUBER (1964), onde foi utilizado um framework assintótico *minimax*. Segundo Peltonen & Kousmanen, Hampel (1974) gerou uma abordagem de estudo de erros em bases de dados baseada em **Funções de Influência** (IF's – *Influence Functions*). A Função de Influência foi introduzida por Hampel sob a denominação de **Curvas de Influência** "como um instrumento heurístico útil de robustez estatística para estudar a performance de estimadores sobre a condição de ruídos".

No trabalho de PELTONEN & KOUSMANEN (2001) a Função de Influencia é utilizada para analisar o resultante de filtros de tamanho finito, seu trabalho se baseia nas curvas de influencia introduzidas por Hampel, para processamento de sinais.

GÜRCAN et. al., (1999) desenvolveu um trabalho utilizando Funções de Influência na área médica, mais precisamente em análise de imagens onde os autores utilizam a notação apresentada em TUKEY(1971) que é derivada da teoria de Hampel, chamada **Curva de Sensibilidade**, que examina o efeito de adicionar um termo *x* isolado na estimação global. Gurcan utiliza este método para analisar a sensibilidade de testes realizados com a equação Gaussiana para verificar quais pontos em uma mamografia podem ser considerados como microcalcificações.

HINNEBURG & KEIM (1998) desenvolveram um algoritmo baseado em densidade, DENCLUE, que utiliza a teoria de funções de influência e a teoria do *Kernel Density Estimation* (KDE) (SILVERMAN, 1986). Este trabalho utiliza a função de influencia para analisar o impacto de cada ponto de dado sobre os seus vizinhos, e somar as influências dos pontos para modelar a densidade global do espaço de dados. A abordagem neste algoritmo foi utilizada como base para o desenvolvimento deste trabalho.

A geração de agrupamentos de pontos de dados utilizando Funções de Influência é estudada através da teoria KDE – *Kernel Density Estimation*. Esta teoria se baseia na idéia de que a influência de cada ponto pode ser modelada utilizando uma função de influência, e que

a densidade global do espaço de dado é calculada pela soma das funções de influência de todos os pontos (HINNEBURG; KEIM, 2003).

A função kernel (K) inicialmente foi desenvolvida para trabalhar com cálculo de densidade de probabilidade. Esta função K satisfaz $\int K(x)dx = 1$. Em notação de escala esta função pode ser descrita:

$$K_h(u) = h^{-1}K\left(\frac{u}{h}\right) \qquad 3.1$$

onde *h* é um número positivo denominado "largura de banda", ou "largura da janela", e *u* unidades de distancia de um ponto a outro (WAND & JONES,1995). Isto permite escrever a função de estimativa *Kernel.*

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \qquad 3.2$$

onde K pode ser uma função de densidade de probabilidade simétrica.

n – é o numero de pontos na vizinhança, numero de balanços sobre o ponto.

X_i - Ponto da vizinhaça observado.

x – Ponto em estimativa.

$f(x)$ - Função de estimativa de densidade.

A função de estimativa *kernel* é a soma dos "balanços" sobre o ponto. A função K estimará a forma destes balanços, baseados na janela *h* que determinará a sua altura, enquanto que a função *f* determinará a sua soma (SILVERMAN, 1986).

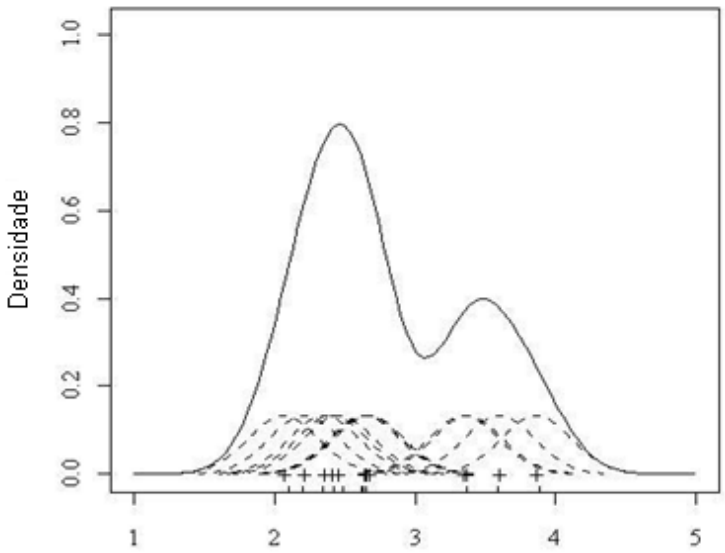

Figura 3.1. Estimativa de densidade *kernel*. Fonte: (WAND & JONES,1995).

A Figura 3.1 ilustra uma função estimativa onde é possível notar que na região em que há muitas observações a densidade é relativamente grande, consequentemente a estimativa *kernel* também irá assumir um valor relativamente grande. O oposto ocorre em regiões onde há relativamente poucas observações (WAND & JONES,1995).

Baseados nestas definições, HINNEBURG & KEIM (2003) desenvolveram uma metodologia para descrever a IF baseada nesta função *kernel*. Nesta notação a contribuição de cada ponto deverá ser calculada separadamente, baseada na distância em que este se encontra de um ponto em estudo.

De acordo com HINNEBURG & KEIM (2003, APUD SILVERMAN, 1986), "a função de *kernel* pode ser vista como uma função que descreve a influencia de um ponto de dados dentro de sua vizinhança". De acordo com uma área de influência limite dada a um ponto é possível verificar quais outros pontos pertencem a sua vizinhança. Com a sua vizinhança definida, é calculada a sua função de densidade, que está relacionada ao cálculo da função *kernel* dos pontos pertencentes à sua vizinhança.

A vizinhança de um objeto de dados em um espaço *d-dimensional* F^d é dada por alguma métrica de distancia apropriada: $F^d x F^d \rightarrow \mathbf{R}$ no espaço F^d. Com isso, a função de densidade de influência de um ponto $x \in F^d$ é definida como a soma das funções de influência de todos os objetos de dados nesse ponto (HINNEBURG & KEIM, 1998).

A partir destas notações as seguintes definições podem ser feitas:

Definição 1. (Função de Influência e Função de Densidade)

Sejam y e x objetos ou pontos em F^d, um espaço *d-dimensional* de entrada. A função de influência de um objeto y em x é uma função $f_B^y : F^d \rightarrow R_0^+$, o qual é definido em termos de uma função de influência básica f^d (HAN & KAMBER, 2006):

$$f_B^y(x) = f^d(x, y) \qquad 3.3$$

A Função de Densidade de um objeto $x \in F^d$ é a soma das funções de influência de todos os pontos de dados. Isto é, a influencia total em x de todos os pontos de dados. Dado *n* objetos de dados descritos por um conjunto de vetores $D \subset F^d$ a **Função de Densidade** é definida como (HAN & KAMBER, 2006):

$$f_B^D(x) = \sum_{i=1}^{n} f_B^{x_i}(x) \qquad 3.4$$

A **Função de Influência** pode ser a principio qualquer função arbitrária, mas é necessário utilizar uma função de distância entre dois pontos pertencentes a mesma vizinhança esta função deve ser reflexiva e simétrica, tal como a distancia Euclidiana (HAN & KAMBER, 2006).

Definição 2. (Atrator de Densidade)

Um ponto $x^* \in F^d$ é um Atrator de Densidade de uma função de influência se x^* é um máximo local de uma função de densidade f_B^D. É desejável que a função de influencia seja, além de simétrica, contínua e diferençável.

Um exemplo de função de influencia básica é a função gaussiana. Nesta função é utilizada uma função de cálculo da distancia entre pontos, assim o valor resultante da gaussiana depende da distancia relativa entre os pontos no espaço de dados em questão (HINNEBURG & KEIM, 1998).

$$f_{Gauss}(x) = e^{-\frac{d(x,y)^2}{2\sigma^2}} \qquad 3.5$$

A função de densidade resultante da função de influência Gaussiana é (HINNEBURG & KEIM, 1998):

$$f_{Gauss}^{D}(x) = \sum_{i=1}^{n} e^{-\frac{d(x,x_i)^2}{2\sigma^2}} \quad 3.6$$

Segundo a definição 2, um Atrator de Densidade de uma função é o máximo local de uma função de densidade. A função densidade tem em 3.6 uma dependência do parâmetro σ. Este parâmetro determina o tamanho da "janela" de influência, ou seja, descreve a influência de um ponto no espaço de dados. Quanto maior o valor de σ, maior é a janela da área de influência, funcionando de forma parecida ao parâmetro h na função kernel. O que quer dizer que pontos dentro desta região sofrem maior influência do que pontos fora desta. Caso o valor deste parâmetro seja alto a influencia de pontos mais distantes também pode aumentar. Apesar disto a posição dos atratores não será afetada diretamente, isto dependerá dos valores de todos os pontos na base.

A Figura 3.2 ilustra três atratores, ou seja, os pontos de máximo da função Gaussiana.

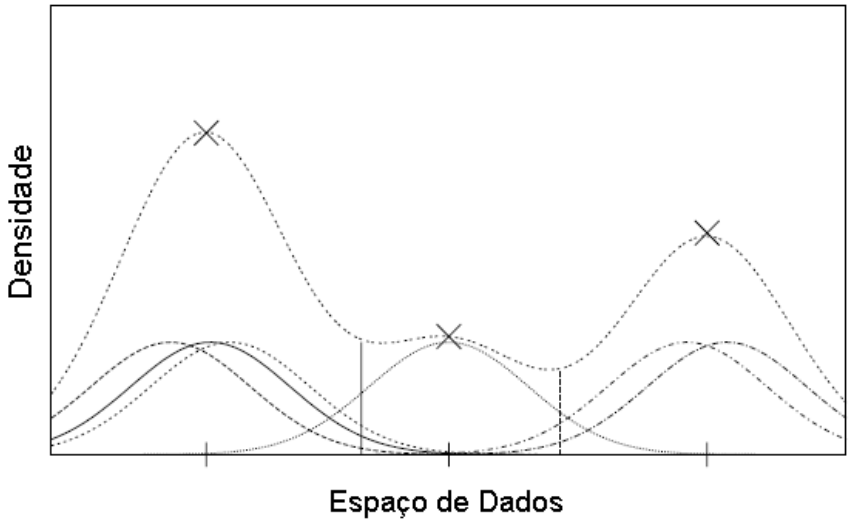

Figura 3.2. Exemplo de atratores de densidade. Fonte: (HINNEBURG & KEIM, 1998)

Os pontos marcados com X representam os atratores de densidade em um espaço de dados unidimensional, a função utilizada é a Gaussiana.

No trabalho de HINNEBURG & KEIM (2003) e HAN & KAMBER (2006) é mencionada a utilização do algoritmo *Hill Climbing* para definir qual ponto é o atrator de densidade. Um atrator de densidade pode ser dito também como um ponto que influência muitos outros.

O parâmetro σ descreve a influência de um ponto em um espaço de dados, é a distância máxima de alcance da influência do ponto x .

4 ALGORITMO GENÉTICO

Segundo FORBES (2005), conforme aumenta a complexidade das tarefas realizadas pelos computadores, pesquisadores buscam inspiração na natureza, que é utilizada como metáfora e modelo, para definir a partir do comportamento e organização de organismos biológicos, novas e melhores maneiras de realizar tais tarefas e futuras outras, assim reinventando a computação.

Uma metodologia desenvolvida com este objetivo foi o Algoritmo Genético (AG).

Os paradigmas do Algoritmo Genético utilizam os processos de seleção ou reprodução, que combinam a seleção dos indivíduos, cruzamento e, posterior, mutação dos indivíduos como o que podemos chamar de **Motor de Otimização**, ou seja, é a força que leva à exploração do espaço de busca de um determinado problema, no sentido de ir atrás de um máximo ou mínimo.

Para poder analisar mais detalhadamente o AG e seu funcionamento, é necessário estudar a sua implementação, e a forma como ele trabalha na busca pela solução do problema. A seguir será apresentado a estrutura de um **Algoritmo Genético**, como este é utilizado comumente para representação das soluções encontradas e a avaliação pela **Função Objetivo** (*fitness*).

4.1 IMPLEMENTAÇÃO DE UM ALGORITMO GENÉTICO

Esta seção apresenta os aspectos fundamentais para a implementação do algoritmo genético. Aqui será apresentado a forma como o problema pode ser representado como um cromossomo (indivíduo), os operadores genéticos que realizam a movimentação do algoritmo no espaço de busca, e a avaliação das soluções encontradas e os parâmetros necessários.

Antes de começar a falar da implementação do AG é necessário definir alguns termos comumente utilizados. São eles, segundo RUTKOWSKI (2008):

- **População:** É um conjunto de indivíduos, normalmente o AG trabalha com um tamanho específico para a população;

- **Indivíduo:** em uma população em Algoritmos Genéticos os indivíduos são conjuntos de parâmetros da tarefa codificados na forma de cromossomos, os indivíduos são as soluções encontradas pelo algoritmo, o que significa dizer que são pontos no espaço de busca. Representados em forma de estruturas de dados.

- **Cromossomo:** É uma seqüência de genes, estrutura que representa o indivíduo.

- **Gene:** Também chamado de uma característica, ou sinal - é um único elemento do genótipo, do cromossomo em particular.

- **Genótipo:** outra estrutura - é um conjunto de cromossomos de um determinado indivíduo. Os indivíduos de uma população podem ser genótipos ou cromossomos simples (caso o genótipo seja formado por cromossomos simples).

- **Fenótipo:** é um conjunto de valores correspondentes a um determinado genótipo, esta é a estrutura decodificada. O fenótipo é o ponto do espaço de busca ao qual o cromossomo representa. É a solução encontrada.

- **Alelo:** é o valor de um determinado gene. É o valor encontrado em uma determinada posição do cromossomo.

- **Lócus:** É o indicativo da posição de um determinado gene na cadeia do cromossomo.

4.1.1 Codificação de problemas

A forma de representar o problema em Algoritmos Genéticos, assim como em qualquer algoritmo de otimização, é muito importante. Diferentes formas de representar o problema levam o algoritmo a trabalhar também de forma diferente. Testes empíricos concluem que no AG a forma de representação pode diferenciar tanto a forma da busca, assim

como a sua qualidade e também o tempo computacional deste é bastante modificado. Um bom estudo na forma de representar pode levar a melhorias drásticas no problema.

Uma característica importante no algoritmo é o **Tamanho da População**. Este é um parâmetro que deve ser informado no início da execução do algoritmo.

Segundo MITCHELL (1999, apud DE JONG, 1975) experimentos realizados por De Jong aplicados a estudos de desempenho em um pequeno conjunto de funções de teste, indicam que uma boa população possui o tamanho no intervalo de 50-100 indivíduos. Conforme MITCHELL, outros estudos também indicam que uma boa configuração do tamanho da população está entre 20-30 indivíduos, independente do problema em teste.

A **População** é a representação das soluções encontradas pelo AG em cada **Geração**. Inicialmente, na primeira geração, a população é criada aleatoriamente, ou seja, o cromossomo dos indivíduos é preenchido com valores aleatórios. E a cada geração, através dos operadores genéticos (que serão apresentados posteriormente), a população evolui e o espaço de busca é estudado.

Outro parâmetro importante para a performance do AG é o **Tamanho do Cromossomo**. Segundo TEIXEIRA (2005) o tamanho do cromossomo está intimamente ligado ao problema abordado e determina o tamanho das soluções candidatas desse problema.

Por último o parâmetro **Número de Gerações.** Este parâmetro determina a quantidade máxima de gerações que serão produzidas pelo AG até este terminar a sua execução, e indicar qual é o indivíduo mais apto, e que conseqüentemente será a solução do problema (TEIXEIRA, 2005).

Outro modo de utilizar o AG é testar caso este atinja uma determinada solução, mas este só é aplicável em casos em que se conheça a solução objetivo.

O desempenho da busca do AG é afetado pelo número de gerações, e tamanho da população, mas a codificação é outro fator significante. O tamanho do cromossomo depende do tipo de codificação. Caso a codificação ajude o desempenho, então a população pode ser aumentada, mas cada caso depende do problema em estudo. A seguir é apresentado alguns tipos de codificação utilizados no AG.

4.1.1.1 Codificação binária

A **Codificação Binária** é a forma comumente utilizada em problemas com AG (TEIXEIRA, 2005). John Holland (HOLLAND, 1975) ao inventar o AG utilizou esta forma de representação.

Esta codificação utiliza uma estrutura de lista para o cromossomo onde cada gene é representado por valores binários, em cada lócus é possível encontrar somente dois alelos possíveis: 0 e 1.

Esta codificação é a mais fácil de trabalhar, já que os operadores genéticos, (mutação, cruzamento e inversão) foram originalmente desenvolvidos para trabalhar com esta codificação.

A sua facilidade de implementação diverge diretamente com a sua capacidade de utilização. Esta implementação é incapaz de ser utilizada em uma série de problemas. Segundo Teixeira ela é propensa a ordenações arbitrarias, o que leva a criação de um número muito grande de soluções inapropriadas, ou até mesmo a não encontrar a solução do problema.

Cromossomo A	101100101100101011100101
Cromossomo B	111111100000110000011111

Figura 4.1 – Exemplo de Codificação Binária das Soluções Candidatas.
Fonte: TEIXEIRA (2005 apud OBITKO,1998)

Ainda existem diversas formas de representar os cromossomos em Algoritmos Genéticos (AG), mas isto depende do problema a ser trabalhado e exige que o AG seja modificado para trabalhar com esta representação.

A seguir serão mostradas formas de como o processo de seleção de indivíduos de uma população pode ser feito para que estes possam realizar o cruzamento entre si, assim movimentar o algoritmo no espaço de busca.

4.1.2 Métodos de seleção

O **Método de Seleção** consiste em selecionar indivíduos baseado em um valor calculado de acordo com a função de *fitness*, para a participação da criação dos "filhos" para a próxima geração (RUTKOWSKI, 2008). Esta seleção ocorre de acordo com a regra de seleção natural concebida por Charles Darwin (DARWIN, 1994), onde os indivíduos mais adaptados de uma população possuem mais chances de gerar descendentes.

Os métodos de seleção escolhem indivíduos da população da geração atual para participarem de uma população transitória, ou temporária, que irá realizar o cruzamento entre seus indivíduos. Esta população é chamada de **População Intermediária**, pois ela se encontra entre os métodos de seleção e os operadores genéticos.

Antes de apresentarmos alguns métodos de seleção, é necessário esclarecer a que se refere quando mencionado a palavra *fitness.*

4.1.2.1 Função de *fitness*

A função de *fitness* é equivalente a concepção de Darwin de adaptabilidade de um indivíduo sobre o ambiente em que vive. De acordo com a sua teoria, com o passar de gerações, indivíduos de uma população, através de mutações e recombinações genéticas, podem adquirir características que influenciam no modo como este se relaciona com o ambiente (DARWIN, 1994).

A função de *fitness* é equivalente a função objetivo de um determinado problema de otimização, e o calculo de seu valor ocorre cada vez que um novo indivíduo é gerado pelo algoritmo.

Matematicamente, tendo Ω como o espaço de busca, a função de *fitness* pode ser dita como (BLASS & MITAVSKIY, 2005):

$f:\Omega \to (0,\infty)$, sendo o objetivo encontrar o valor máximo da função de *fitness f*.

Uma população *P* do AG é representada como a seguir:

$$P = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad 4.1$$

Onde, x_i representa uma solução encontrada, indivíduo da população, pertencente ao espaço de busca Ω . A função de *fitness* avalia cada indivíduo da população.

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \begin{matrix} \rightarrow [f(x)]_1 \\ \rightarrow f(x_2) \\ \vdots \\ \rightarrow f(x_n) \end{matrix} \quad 4.2$$

Com o valor de fitness calculado é possível aplicar algum método de seleção. Existem muitos deles, um dos mais utilizados é o método de seleção chamado roleta. Alguns destes serão apresentados a seguir.

4.1.2.2 Seleção por roleta

Este método de seleção "imita" uma **Roleta**, a mesma utilizada em jogos de azar. Segundo TEIXEIRA (2005) este método é o mais comumente utilizado nas implementações de algoritmos genéticos. Este é o método de seleção original proposto por Holland em seu algoritmo genético (HOLLAND, 1975). Neste método a probabilidade de selecionar um certo indivíduo da população é proporcional a sua aptidão, *fitness*.

A probabilidade de seleção $r(x_i)$ é calculada como apresentado a seguir (BLASS & MITAVSKIY, 2005):

$$r_i = \frac{f(x_i)}{\sum_{i=1}^{n} f(x_i)} \quad 4.3$$

A roleta pode ser construída dividindo-a em setores correspondentes a cada indivíduo x_i, e de tamanhos determinados de acordo com a probabilidade de seleção de cada indivíduo (r_i). Sendo assim, quando a roleta for girada, a probabilidade de parada em x_i é r_i (TEIXEIRA, 2005 apud BORGES, 2002).

Cada vez que a roleta é "girada" um indivíduo é selecionado para participar da população intermediária, este processo pode ser repetido até o número de indivíduos da população intermediaria alcançar o desejado. Quanto maior a área de um cromossomo na roleta, maior a probabilidade deste ser selecionado, conseqüentemente mais vezes este pode ser selecionado para cruzamento, produzindo assim uma maior quantidade de descendentes.

A tabela 4.1 é um exemplo de uma população de 4 indivíduos de um AG, como pode ser visto cada um deles possui uma fatia da roleta proporcional ao seu *fitness* e igual ao percentual de seu *fitness,* que foi calculado de acordo com a sua probabilidade de seleção. Por exemplo, o indivíduo um (1) possui uma probabilidade de seleção igual a 0,144.

Tabela 4.1 – Valores de aptidão calculados para cada indivíduo e a proporção da roleta

N°.	Indivíduos	(*Fitness*)	% do Total
1	01101	169	14,4
2	11000	576	49,2
3	01000	64	5,5
4	10011	361	30,9
	Total	1170	100,0

Fonte: TEIXEIRA (2005, apud GOLDBERG, 1989)

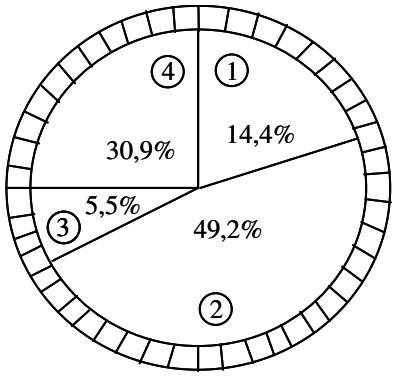

Figura 4.4 – Roleta dividida entre os indivíduos
Fonte: TEIXEIRA (2005, apud GOLDBERG, 1989)

4.1.2.3 Seleção por *rank*

A noção de seleção por ***Rank*** foi introduzida por BAKER (1985) como uma alternativa ao método de seleção proporcional ao *fitness*, sendo um método semelhante. Segundo TEIXEIRA (2005), tem como objetivo prevenir uma tendência rápida de convergência a um ótimo local. Sendo assim busca equilibrar a probabilidade de seleção dos indivíduos, mas mantendo a característica de que os indivíduos mais adaptados, com melhor *fitness*, possuem maior probabilidade de seleção.

Ele funciona da seguinte forma: primeiramente a população deve ser ordenada do melhor ao pior, de acordo com o seu valor de *fitness*. Cada indivíduo deve receber um novo valor, de acordo com a sua posição na ordenação. Dados *N* indivíduos em uma população, a posição *N* deve ser dada ao melhor indivíduo, e a posição 1 ao pior (GOLDBERG & DEB, 1991). Após isso pode ser realizada uma seleção proporcional a este valor dos indivíduos que participarão da População Intermediária. Este processo pode ser repetido até que o número de indivíduos da população intermediaria alcançe o desejado.

A probabilidade de seleção do indivíduo se dará de acordo com a sua posição no *rank*. De acordo com (BLICKLE & THIELE, 1995) deve ser notado neste método que mesmo que os indivíduos possuam o mesmo valor de *fitness*, estes possuirão diferentes posições no *rank*, sendo assim, possuirão diferentes probabilidades de seleção.

No tópico seguinte serão apresentados os **Operadores Genéticos**. São operadores que realizam o processo de expansão do espaço de busca pelo algoritmo, através da variabilidade genética da população. São eles o cruzamento, inversão e a mutação.

4.1.3 Operadores genéticos

Os operadores genéticos determinam a produção de sucessores de um AG. É compreendido por um conjunto de operadores que recombinam e mutam membros selecionados da população atual (MITCHELL, 1997). Estes operadores correspondem a versão idealizada dos operadores genéticos encontrado na evolução biológica.

Os operadores genéticos são relacionados a seguir.

4.1.3.1 Cruzamento

O operador de cruzamento produz dois novos descendentes de um par de indivíduos. Denominados “pais”, pela cópia de determinados genes de seus cromossomos.

Conforme MITCHELL (1999) este operador é a principal fonte do poder do AG, com a habilidade de recombinar instancias de cromossomos de forma a produzir instâncias igualmente boas ou melhores. O operador imita grosseiramente a recombinação biológica entre dois organismos de cromossomos simples (haplóide).

Um parâmetro importante a ser informado para o cruzamento é a **Taxa de Cruzamento.** A Taxa de Cruzamento implica na geração de descendentes. Esta é uma probabilidade de haver cruzamento em um determinado par de “pais”.

Segundo MITCHELL (1999) estudos realizados apontam que uma boa taxa é de ~0.6 por par de pais escolhidos para realizar cruzamento, mas Mitchell também aponta estudos em que esta taxa é boa com valor entre 0.75 e 0.95.

Existem diversos tipos de cruzamento, a seguir será apresentado o cruzamento **Uniforme**, método utilizado neste trabalho

4.1.3.1.1 Cruzamento uniforme

O cruzamento uniforme é realizado pela geração de padrões de 0s e 1s estocasticamente, utilizando a distribuição de Bernoulli, funcionando como uma “mascara” binária (WHITLEY, 1993).

A distribuição de Bernoulli é utilizada em no espaço amostral discreto [0,1], onde a probabilidade $P(0) = 1 - p$ e $P(1) = p$. Esta idéia foi primeiramente utilizada por SYSWERDA (1989), que assumiu implicitamente o parâmetro de Bernoulli $p = 0.5$. Mas este pode ser modificado de acordo com as necessidades (WHITLEY, 1993).

Conforme TEIXEIRA (2005) neste cruzamento cada gene do descendente é gerado com base na cópia do gene correspondente em um dos pais, de acordo com uma “máscara”, gerada aleatoriamente, de tamanho igual ao cromossomo. Caso o conteúdo de uma respectiva posição da máscara seja igual a um (1), então o descendente recebe o gene equivalente a mesma posição do primeiro pai, caso contrário, ou seja, o conteúdo da posição

igual a zero (0), então o gene a ser copiado será do segundo pai. A figura 4.7 ilustra o que foi dito.

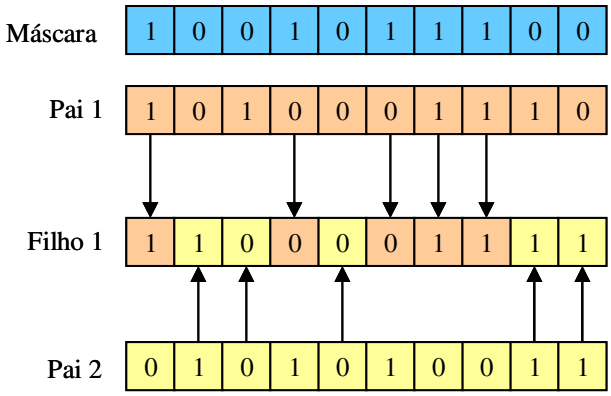

Figura 4.7 – Exemplificação da operação de cruzamento uniforme. Fonte: TEIXEIRA (2005, apud BORGES (2002)

Este processo é repetido para todos os cruzamentos realizados pela **População Intermediária**. Da mesma forma, a cada nova geração uma nova mascara deve ser construída. Com este processo os filhos irão conter uma mistura dos genes dos pais.

O processo de mutação garante a variabilidade genética da população assim como a abrangência no espaço de busca.

4.1.3.2 Mutação

O operador de mutação tem o poder de possibilitar que todos os cromossomos possíveis sejam alcançados (FANG, 1994).

A mutação funciona como um operador que ocasionalmente modifica o valor de um determinado gene e permite que alelos alternativos sejam revistados (WHITLEY, 1993).

Conforme MITCHELL (1999) há estudos em que uma boa taxa de mutação utiliza o valor de 0.001 por cada gene do cromossomo. Ou seja, com este estudo cada gene será testado de acordo com esta taxa, de 1/1000. Mas Mitchell também aponta estudos em que esta taxa está entre 0.005 e 0.01 por indivíduo da população.

Caso esta taxa seja mais elevada, o algoritmo poderá passar a funcionar como uma busca aleatória, devido a alta taxa de modificação dos genes.

Na figura 4.9 é exemplificada a operação de mutação em um cromossomo composto de treze genes. Nestes, o gene sete (7) sofre a mutação. O valor deste gene que era zero (0), após a mutação, passará a ser um (1).

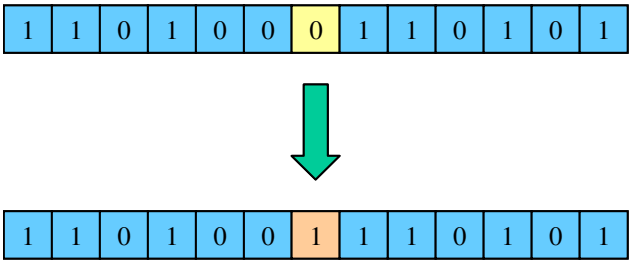

Figura 4.9 – Exemplificação da operação de mutação

Fonte: TEIXEIRA (2005 apud BORGES (2002)

Alguns autores aplicam de formas diferentes o método de mutação, REEVES & ROWE (2003 apud DE JONG, 1975) utilizam uma distribuição de Bernoulli em cada lócus com uma baixa probabilidade da mutação ocorrer, ocasionando algo parecido com uma "mascara" binária onde, caso o valor seja 1, a mutação ocorre e, caso seja 0, a mutação não ocorre.

No capítulo seguinte será explicado o algoritmo de clusterização baseado em densidade e grade desenvolvido.

5 ALGORITMO GENÉTICO PARA CLUSTERIZAÇÃO BASEADO EM DENSIDADE E GRADE

Durante a produção deste trabalho foi desenvolvido um algoritmo para agrupamento utilizando Algoritmo Genético (AG) e baseado nas metodologias de Densidade e Grade de agrupamentos mencionados nas seções 2.1.2 e 2.1.3. Mais especificamente esta metodologia se baseia no algoritmo EDACluster (OLIVEIRA, 2007) que é um algoritmo evolutivo para análise de agrupamento, também baseado em densidade e grade, que utiliza um método de estimativa de distribuição no algoritmo evolutivo, ao invés de utilizar os tradicionais métodos de cruzamento e mutação.

No método proposto, ao invés de utilizar os próprios objetos da base de dados, o agrupamento ocorre através da seleção de células centrais para cada agrupamento, e a partir daí as outras células são adicionadas de forma iterativa, respeitando a condição de que as células são adicionadas somente ao grupo referente à célula centro mais próxima. Assim, antes deste agrupamento, ocorre uma etapa de configuração, onde as células serão criadas. O agrupamento de células reduz a complexidade do problema, já que o número de células pode ser muitas vezes menor do que o número de objetos da base.

A grade é formada pela divisão do espaço de *features* em células. Cada objeto irá pertencer a célula correspondente ao seu valor em cada atributo. A divisão do espaço em células possibilita agrupar um subespaço do espaço de *features* nos quais os objetos estarão contidos, assim como o espaço continuo “vazio”, onde não há objetos. Futuramente, este espaço continuo será útil na aplicação do algoritmo de cálculo de pontos atratores.

A densidade é utilizada como critério de formação dos agrupamentos, além da distância das células. O agrupamento de densidade é utilizado para agrupar a base de dados de acordo com a densidade das células, formando assim grupos de alta densidade.

Outro critério utilizado para a formação dos grupos é a minimização da distância de uma célula a uma célula central. Desta forma é possível formar grupos, além de alta densidade, adicionando às células mais próximas. Com isso busca-se agrupar as células de acordo com a sua proximidade, além de agrupar de acordo com a densidade do agrupamento,

particionando o espaço de *features* de forma que o agrupamento valorize a densidade das células com o mínimo de células possíveis.

Este capítulo descreve as etapas necessárias para a utilização deste algoritmo, desde a etapa de configuração até a função de avaliação utilizada no AG.

5.1 CONFIGURAÇÃO DA BASE

Primeiramente é necessário que a base de dados a ser utilizado passe por uma etapa de configuração. Nesta etapa a base passará a ser representada por uma grade de *d* dimensões, onde cada dimensão é um atributo dos dados.

Cada atributo deverá ser dividido em um número de faixas. Estes deverão ser igualmente espaçados, dessa forma os objetos passarão a ser representados pelas células nos quais estarão contidos seus respectivos valores de atributos. Sendo assim, como as células possuem o mesmo volume, a densidade da célula é definida pelo número de itens nela contidos (OLIVEIRA, 2007).

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

A Figura 5.1 apresenta um exemplo da etapa de configuração de uma base de dados bidimensional. No esquema de configuração proposto, os atributos Salário e Idade foram divididos em 6 (seis) faixas de valores. Os objetos que estão entre os valores 30-35 do atributo idade e 300-600 para o atributo salário passarão a ser representados pela célula (0,2) cuja densidade é igual a 2 (dois).

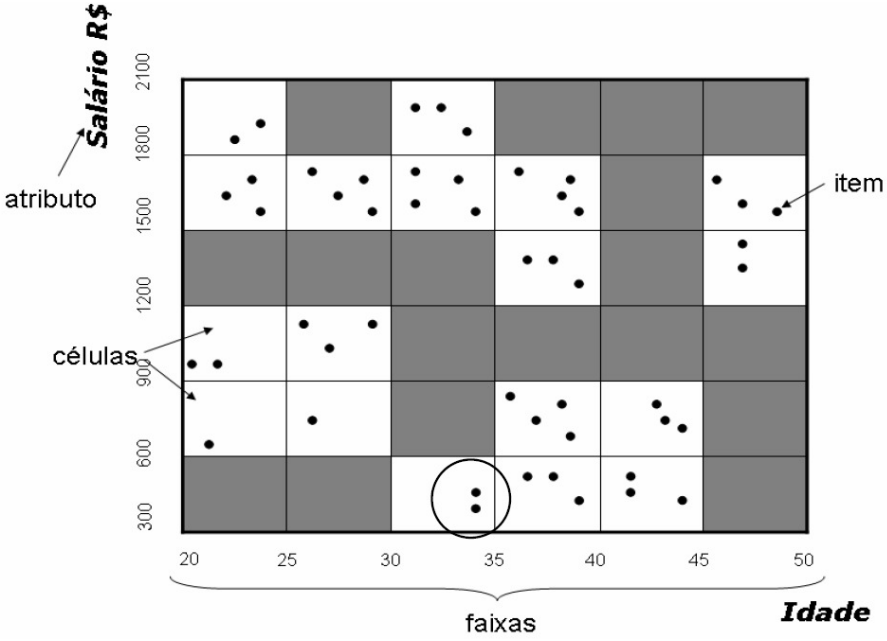

Figura 5.1 - Representação do plano da base em células de uma grade.

Fonte: OLIVEIRA (2007)

Uma comparação pode ser feita com a utilização desta etapa de configuração em um algoritmo de agrupamento por particionamento. O algoritmo *K-Means*, por exemplo, possui complexidade computacional igual a $O(n \cdot k \cdot t)$, onde n é o número total de objetos, k é o número de clusters, e t é o número de iterações (HAN & KAMBER, 2006). Para o exemplo apresentado na figura 5.1 o valor de n é igual a 51. Aplicando a etapa de configuração e eliminando as células que não possuem objetos acabam por restar somente 18 células, ou seja, menos da metade do trabalho computacional necessária ao agrupamento do problema original.

De acordo com o apresentado, a ordem da complexidade continua sendo a mesma, mas o trabalho computacional diminui bastante, tornando menos onerosa a tarefa de agrupamento.

A seguir será apresentado a forma como o Algoritmo Genético foi utilizado neste método.

5.2 REPRESENTAÇÃO DO CROMOSSOMO

O objetivo do AG neste trabalho é encontrar as células mais adequadas para serem os centros de cada agrupamento. Com isso, o cromossomo apresenta as posições destas células centrais encontradas.

Para fins de melhorar o desempenho do algoritmo foi utilizado o cromossomo binário apresentado na seção 4.1.1.1 por facilitar a implementação do método de cruzamento e mutação, consequentemente tornando o desempenho melhor do que se aplicado outras formas de representação.

O cromossomo utilizado neste trabalho possui um tamanho *t*, onde este tamanho depende do número de faixas (*f*) utilizadas na etapa de configuração e quantos *bits* são necessários para representar as faixas. Se as faixas começarem a partir do 0, então deverão possuir valores entre 0 e $f - 1$. $bits(f-1)$ representa o número de bits para representar as faixas. O tamanho do cromossomo também dependerá do número de atributos (*a*) da base (dimensões), e o número de agrupamentos (*g*) que se deseja encontrar. Ou seja

$$t = bits(f-1) \cdot a \cdot g \tag{5.1}$$

A Figura 5.3 apresenta um exemplo de cromossomo utilizado para representar as células escolhidas na figura 5.2.

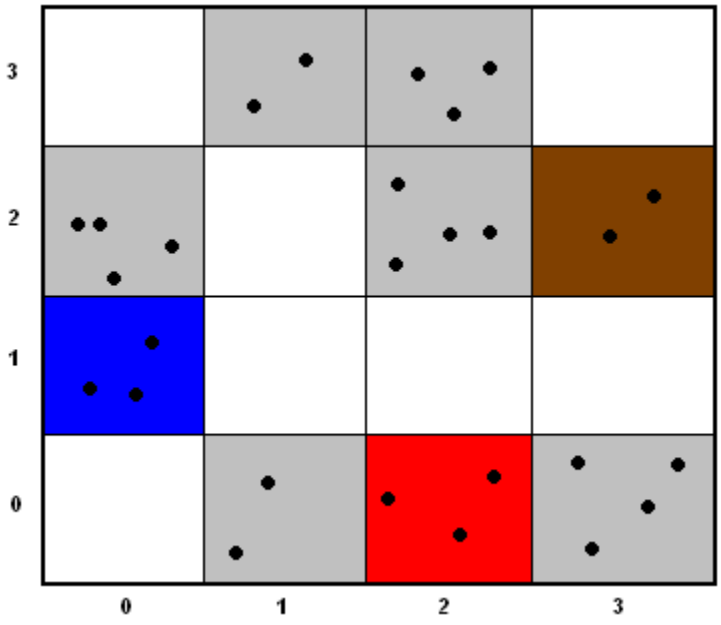

Figura 5.2. Exemplo de base bidimensional com as 3 células centrais escolhidas.

De acordo com a figura o número de faixas *f* em que a base foi dividida é igual a 4 (0-3), o número de agrupamentos *g*, ou seja, o número de células centrais, é igual a 3, e o número de atributos *a* da base é igual a 2. Convertendo o numero de faixas *f* = 4 para binários *bits(4-1)* temos que o número de *bits* para representar esse valor é igual a 2. Estes valores requisitam um cromossomo de tamanho *t* = 12, conforme ilustrado na Figura 5.3

0 1 0 0 0 0 1 0 1 0 1 1

Figura 5.3. Exemplo de cromossomo gerado pelo Algoritmo Genético

A cada quatro (4) alelos deste cromossomo obtemos as posições de uma célula centro. E a cada dois (2) destes obtemos o valor de uma faixa de um determinado atributo, como pode ser visto na figura a seguir.

CENTRO 1		CENTRO 2		CENTRO 3	
Atributo x	Atributo y	Atributo x	Atributo y	Atributo x	Atributo y
01	00	00	10	10	11

Figura 5.4. Cromossomo dividido em centros e faixas.

Obtendo os valores binários que representam uma faixa, é possível utilizar a conversão de valores binários para valores decimais para representar o valor da faixa escolhida.

Por exemplo, considerando uma seqüência de *bits B*, $[a_B, \ldots, a_2, a_1, a_0]$, o valor decimal correspondente pode ser obtido calculando-se (SILVA 2006):

$$\overline{x} = \sum_{i=0}^{m-1} a_i \cdot 2^i \qquad 5.2$$

e, em seguida:

$$x = a + \overline{x}\left(\frac{b-a}{2^m - 1}\right) \qquad 5.3$$

onde:

a e *b* definem o domínio de variação dos valores da variável x e

m representa o comprimento total do gene.

Por exemplo, considerando a sequência B = [0,1], que representam a primeira faixa, convertendo para base 10

$$\overline{x} = \sum_{i=0}^{m-1} a_i \cdot 2^i = 0 \cdot 2^1 + 1 \cdot 2^0 = 1$$

e, considerando que o valor das faixas está entre 0-3, e que o numero de bits utilizado na representação é igual a 1, o valor de x é:

$$x = 0 + 1\left(\frac{3-0}{2^1-1}\right) = 1$$

ou seja, o valor da primeira faixa.

A Figura 5.5 apresenta os valores de todas as faixas do exemplo na Figura 5.4 convertidos para decimal.

CENTRO 1		CENTRO 2		CENTRO 3	
Atributo x	Atributo y	Atributo x	Atributo y	Atributo x	Atributo y
1	0	0	2	2	3

Figura 5.5. Cromossomo com seus genes convertidos de valores binários para valores decimais.

Obtendo as posições das células centrais escolhidas após o processo de conversão, é possível aplicar o processo de formação do agrupamento, no tópico 5.4.

5.3 POPULAÇÃO INICIAL

A população é gerada aleatoriamente com um número definido de indivíduos. Como pôde ser visto na seção 4.1.1 uma boa população inicial pode estar no intervalo entre 20-100 indivíduos.

O cromossomo de cada indivíduo é preenchido com valores binários aleatoriamente, alelo a alelo, até o cromossomo ser completado.

Um dos problemas que podem surgir é com relação à geração de um cromossomo que possua uma célula-centro em sua codificação mais de uma vez. Este problema deve ser tratado no momento em que há a formação do agrupamento, e a sua avaliação. Neste trabalho é fornecido um valor extremamente baixo de *fitness* caso isso ocorra.

5.4 FORMAÇÃO DO AGRUPAMENTO

Para a formação dos grupos é utilizado um método que verifica as distâncias das células da grade, as que contêm objetos, a cada um dos centros escolhidos (genes do cromossomo). Caso uma célula c_i esteja mais próxima de um dado cento c_k então esta célula é atribuída ao grupo C_k correspondente a este centro, ou seja:

$$\|c_i - c_k\| < \|c_i - c_p\|, \; p=1,2,...,K \text{ e } k \neq p \; .$$

A distância de Spearman Footrule, também chamada de Distancia de Manhattan, (TEKNOMO, 2010; ALVO & PAN, 1997) é utilizada para calcular a distância entre as células ocupadas da grade multidimensional e os centros candidatos dos grupos.

$$d_{ik} = \sum_{l=1}^{a} |c_{il} - c_{kl}| \qquad 5.4$$

onde:

d_{ik} é a distância da célula c_i para o centro c_k;

a é o número de dimensões, atributos, da grade;

l é o índice do atributo para célula atual;

c_{il} é a posição da célula c_i no atributo l;

c_{kl} é a posição da célula-centro c_k no atributo l;

Por exemplo, na Figura 5.2, a distância da célula (3,1) para a célula-centro (2,3) em vermelho:

$$d_{ik} = |3-2|+|1-3| \therefore \; d_{ik} = 1+2 \therefore \; d_{ij} = 3$$

Considerando a distância igual a 3 (três), se a célula (3,1) estiver mais próxima da célula (2,3), do que das outras duas células centros, então (3,1) será adicionada ao agrupamento pertencente a célula-centro (2,3) e a sua densidade será atualizada com a densidade de (3,1).

5.5 FUNÇÃO DE AVALIAÇÃO

Após os grupos serem formados, é possível avalia-los baseado em suas densidades e a média das distancias entre as células que formam o grupo. A função de avaliação é inspirada no algoritmo de agrupamento CLIQUE (AGRAWAL et al., 1998) adicionando-se o critério de minimização das distâncias entre as células dos grupos.

A densidade é calculada pela razão entre o número de itens de um grupo e o número de células do grupo. Após o cálculo da densidade, realizado no momento da formação do cluster, é possível calcular o *fitness* de um indivíduo levando-se em consideração a média das densidades como em Oliveira (2007) e a média das distâncias das células pertencentes ao grupo e as respectivas células centrais, ou seja:

$$densidade(k) = \frac{numeroItens(k)}{m_k} \tag{5.5}$$

$$mediaDistancia(k) = \frac{\sum_{i=1}^{m_k} d_{ik}}{m_k} \tag{5.6}$$

$$fitness_i = \prod_{k=1}^{K} \frac{densidade(k)}{mediaDistancia(k)} \tag{5.7}$$

onde,

K é o número de grupos, consequentemente, número de células centrais.

m_k é o número de células no grupo *k*.

d_{ik} é a distancia da célula *i* para a célula centro *k*.

Assim, a *fitness* de um indivíduo é proporcional ao produto das densidades dos agrupamentos e inversamente proporcional ao produto da média das distâncias das células nestes. O que quer dizer que quanto maior as densidades dos agrupamentos, ou seja, menor número de células com grande quantidade de objetos, maior será o *fitness*. Além do mais, quanto menor as distâncias entre as células em cada um dos agrupamentos melhor será o indivíduo e maior a sua probabilidade de passar as suas características para as futuras gerações.

5.6 MÉTODO DE SELEÇÃO E OPERADORES GENÉTICOS

Neste trabalho foram utilizados os operadores genéticos comuns do AG, Seleção, Cruzamento e Mutação. Estes operadores foram explicados no capítulo 4 deste trabalho. Os tópicos a seguir apresentam a aplicação destes operadores no cromossomo definido anteriormente.

5.6.1 Seleção

O método de seleção é aplicado para selecionar os indivíduos com melhor valor de *fitness* da população para participarem do processo de cruzamento. Um número de indivíduos a serem escolhidos deve ser definido. Estes indivíduos selecionados são chamados de população Intermediária.

O método de seleção utilizado é uma forma híbrida da seleção por *Rank* com o método de Roleta. Primeiramente é feito um *ranking* dos indivíduos da população e após isso o método da roleta é aplicado para selecionar os indivíduos que participarão do cruzamento.

O *ranking* ordena os indivíduos da população de forma que os indivíduos que possuem melhor valor de *fitness* fiquem nas ultimas posições. Esta ordenação pode ser realizada por qualquer algoritmo de ordenação, mas para este trabalho foi utilizado o método de *Quicksort* (HOARE, 1962). A seleção utilizando o método de *Rank* não utiliza o seu valor de *fitness*, mas sim a sua posição no *Rank*, o que quer dizer que quanto maior a sua posição, maior a sua probabilidade de seleção. Por exemplo, se a população possui tamanho igual a 10, de acordo com o *Rank* o melhor individuo deverá estar na posição 10 e, o seu valor deverá ser

10, enquanto que o pior, que está na posição 1, deverá ter valor 1. Mais detalhes do método de *Rank* na seção 4.1.2.3.

Após a ordenação é aplicado o método de Roleta para a seleção, mas agora ao invés de utilizar a proporcionalidade ao *fitness,* deverá ser utilizada a proporcionalidade ao valor no *Rank*. No exemplo anterior, o melhor indivíduo, na posição 10, possui valor 10, a sua probabilidade de seleção r_i é feita da seguinte forma:

$$r_i = \frac{R(x_i)}{\sum_{i=1}^{n} R(x_i)} \qquad 5.8$$

Onde $R(x_i)$ é o valor no Rank do indivíduo x_i.

Com isso a probabilidade de Seleção do indivíduo anterior é $r_1 = \frac{10}{55}$, ou $r_i =$ 0.182, enquanto o indivíduo na posição 1 possui probabilidade $r_{10} = \frac{1}{55}$, ou $r_1 = 0.018$. Tendo as probabilidades, a roleta abstrata pode ser formada e o método aplicado para a seleção dos indivíduos da subpopulação. Mais detalhes sobre o método da roleta na seção 4.1.2.2. A seguir o método de cruzamento utilizado nesta subpopulação escolhida.

5.6.2 Cruzamento

O método de cruzamento tem o objetivo de recombinar os cromossomos dos indivíduos selecionados pelo método de seleção, apresentado anteriormente, a fim de gerar novos cromossomos potencialmente bons para a nova geração.

O método de cruzamento utilizado neste trabalho foi o Cruzamento Uniforme. Uma máscara formada por valores binários é criada inicialmente de forma aleatória, com o mesmo tamanho do cromossomo dos indivíduos. A cada vez que um novo cruzamento é realizado, e utilizando desta mascara, é escolhido qual valor de gene será armazenado, se o do primeiro pai ou o do segundo pai, visto que em uma máscara binária somente dois pais podem ser utilizados. A probabilidade de ser realizado cruzamento, ou seja, a taxa de cruzamento escolhida é 0.8. Mais detalhes sobre este método de cruzamento na seção 4.1.3.1.1.

5.6.3 Mutação

O método de mutação tem o objetivo de expandir o espaço de busca no Algoritmo Genético. Este método modifica um valor de gene substituindo o valor atual por algum completamente novo, a fim de inserir novos dados na população.

Como o cromossomo utilizado possui somente valores binários, então o método de mutação somente modifica um valor de gene de 0 para 1 ou vice-versa.

Primeiramente, a cada vez que uma mutação será realizada, uma máscara de valores binários é gerada de modo aleatório com o mesmo tamanho do cromossomo do indivíduo a ser mudado. A mascara é então percorrida e onde for encontrado o valor 1, significa que o gene na mesma posição, no indivíduo em questão, deve ser modificado.

A probabilidade de mutação utilizada foi de 0.01, o que quer dizer que existe 1% de probabilidade de um gene aparecer com valor 1 na mascara. Mais detalhes sobre o método de mutação na seção 4.1.3.3.

5.7 PSEUDOCÓDIGO DO ALGORITMO

A Figura 5.6 apresenta o algoritmo genético baseado em densidade e grade para o agrupamento dos dados desenvolvido neste trabalho.

Primeiramente o algoritmo começa com a **configuração**, onde o espaço de dados será configurado e dividido em f faixas de forma que fique parecido com uma grade multidimensional, formado por espaços menores (células), com intervalo igual pela intersecção das faixas. Após esse passo a população é iniciada com a criação dos indivíduos que dela participam. Cada indivíduo possui um cromossomo de tamanho t, calculado de acordo com a Expressão (5.1), de valores binários. O cromossomo representa células da grade que foi escolhida para ser um centro de um agrupamento.

O Algoritmo Genético então avalia a população gerada em cada geração, por um determinado número de gerações pré-definidas. A avaliação gera os agrupamentos, conforme a Função (5.4), para verificar a viabilidade das células centrais escolhidas, calculando assim o *fitness* dos indivíduos através das expressões (5.5), (5.6) e (5.7).

```
Algoritmo de Agrupamento Baseado em Densidade e Grade Utilizando Algoritmo Genético
1   //Fase de configuração
2   Para cada atributo da base faça
3       Encontrar valor máximo e mínimo
4       Encontrar o intervalo das faixas.
5       Dividir o atributo em faixas
6   Fim do Para
7   Para cada objeto da base faça
8       Para cada célula da grade faça
9           Verificar se o objeto pertence ao intervalo dos atributos da célula
10          Caso pertença adicionar indicador da célula ao atributo
11      Fim do Para
12  Fim do Para
13  //Algoritmo Genético
14  Inicializar População Inicial
15  Para cada indivíduo faça
16      Iniciar cromossomo aleatoriamente
17      Avaliar cromossomo com construir_Agrupamentos(cromossomo)
18      Calcular fitness cromossomo pela formula 5.7
19  Fim do Para
20  Para cada geração faca
21      Selecionar os indivíduos de acordo com o método de seleção
22      Realizar cruzamento de acordo com probabilidade de cruzamento
23      Gerar descendentes
24      Para cada Individuo da nova_geração faça
25          Realizar Mutação de acordo com probabilidade de mutação
26          Avaliar cromossomo com construir_Agrupamentos(cromossomo)
27          Calcular fitness cromossomo pela formula 5.7
28      Fim do Para
29      Inserir descendentes na população
30      Armazenar melhor indivíduo da população
31  Fim do Para
32  //Construção do agrupamento
33  Procedimento construir_Agrupamento(cromossomo)
34      Dividir cromossomo de acordo com o número de grupos
35      Converter cromossomo binário Para decimal
36      Para cada centro x do cromossomo faça
37          Criar agrupamento inicial com este centro
38          Calcular densidade(grupo x) pela formula 5.5
39          Calcular mediaDistancia(grupo x) pela formula 5.6
40      Fim do Para
41      Para cada célula da grade faça
42          Para cada centro x do cromossomo faça
43              Verificar a distancia da célula Para o centro
44              Se a distancia  Para centro x menor do que Para qualquer outro centro então
45                  Adiciona célula Para grupo x
46                  Calcular nova densidade(grupo x)
47                  Calcular nova mediaDistancia(grupo x)
48              Fim do então
49          Fim do Para
50      Fim do Para
51  Fim Procedimento
52  Retorna Agrupamentos Gerados
```

Figura 5.6. Algoritmo de Agrupamento Baseado em Densidade e Grade Utilizando Algoritmo Genético (AGABDG)

Esta população passa por um processo de seleção, através do método definido na seção 5.6.1, e a população intermediária selecionada pelo processo de cruzamento e mutação, definidos nas Seções 5.6.2 e 5.6.3.

O algoritmo então armazena o melhor indivíduo em cada geração, para que as suas características possam ser transmitidas para as gerações futuras. E o procedimento termina com a geração de um arquivo com os novos agrupamentos formados pelo melhor indivíduo encontrado pelo AGABDG em todas as gerações.

6 CÁLCULO DE PONTOS DE ALTA DENSIDADE EM GRUPOS DE DADOS

O algoritmo de agrupamento proposto no Capitulo 5 realiza uma busca na base de dados e organiza os mesmos em grupos. Estes agrupamentos são conjuntos formados pelos objetos mais similares, com valores de seus atributos mais próximos.

Tendo em mãos estes grupos, o algoritmo de cálculo de densidade utilizando Funções de Influência pode ser aplicado para descobrir em cada agrupamento os pontos na base considerados atratores de densidade.

No contexto desta dissertação, atratores de densidade são pontos máximos da função densidade de influência nas regiões (células) do espaço de dados que contém cada um dos agrupamentos gerados pelo algoritmo descrito no Capítulo 5. São pontos da base que possuem um grande conjunto de dados em sua vizinhança, de acordo com um limiar σ de distancia, o que gera alta densidade devido à influência que sofre dos pontos vizinhos.

Para este método os agrupamentos já são conhecidos, com isso é possível encontrar os pontos que atraem estes agrupamentos. Para isso basta encontrar os pontos atratores de cada região que contém o agrupamento. Cada região formada pelas células que contém o agrupamento é tratada com um AG. Pode-se dizer que a região que contem as células do agrupamento, incluindo células vizinhas "vazias", será utilizada no cálculo dos pontos atratores. E os atratores serão encontrados separadamente. Este trabalho apresenta o algoritmo genético desenvolvido para o cálculo dos pontos atratores.

6.1 REPRESENTAÇÃO DO CROMOSSOMO

A representação do cromossomo no algoritmo também é a binária, onde os genes são representados por bits (0,1). O tamanho *t* do cromossomo deve ser calculado considerando o domínio do espaço definido pelas células que contém os agrupamentos gerados pelo AG apresentado no capítulo anterior. Por exemplo, considerando um atributo cujos valores estão dentro do intervalo [0.00, 9.99], existem pelo menos 1000 possibilidades de valores distintos para este atributo. Em binários 10 *bits* são suficientes para representar estes valores, logo o comprimento do cromossomo pode ser obtido:

$$t = \sum_{i=1}^{n} bits(a_i) \qquad 6.1$$

onde *n* é o numero de atributos/dimensões de cada instância na base, *bits*(a_i) é o número mínimo de *bits* necessários para representar o domínio de um atributo i

Considerando este esquema de representação adotado para o cromossomo, a população inicial pode ser gerada de maneira semelhante à descrita na Seção 5.3. Cada indivíduo a ser gerado o cromossomo é preenchido com uma sequência aleatória de *bits*, calculado de acordo com a Expressão (6.1).

6.2 FUNÇÃO DE AVALIAÇÃO

Considerando a população inicial gerada, a função de avaliação calcula o *fitness* de seus indivíduos. Os valores binários do cromossomo devem ser convertidos para a base 10 e normalizados de acordo com os valores máximos e mínimos de cada atributo. Esta conversão pode ser realizada empregando-se as Expressões (5.2) e (5.3).

Tendo o cromossomo sido convertido para valores na base 10 é possível calcular o valor do *fitness* deste cromossomo. Neste trabalho o Algoritmo Genético busca um ponto no espaço formado pelas células que tenha máxima densidade, segundo a função de densidade (3.6).

Considerando $D = (x_1, \ldots, x_n)$ os pontos pertencentes ao agrupamento em que a busca está sendo realizada, e *x* o ponto no espaço de dados encontrado pelo Algoritmo Genético, o AG irá buscar um ponto *x* em que a função densidade seja máxima:

$$fitness^{D}(x) = \sum_{i=1}^{n} e^{-\frac{d(x,x_i)^2}{2\sigma^2}} \qquad 6.2$$

Com o valor do *fitness* do indivíduo calculado, o método de seleção pode ser aplicado para selecionar os indivíduos que irão participar da etapa de cruzamento. Ambos são expostos a seguir.

6.3 MÉTODO DE SELEÇÃO E OPERADORES GENÉTICOS

O método de seleção aplicado neste algoritmo para encontrar os atratores foi o mesmo utilizado no processo de seleção dos indivíduos no algoritmo de agrupamento no Capítulo 5. Primeiramente um ranking da população é realizado considerando o *fitness* dos indivíduos pelo método do *Quicksort*. Em seguida, é aplicado o método da Roleta com a seleção proporcional a sua posição do *rank*.

O método de cruzamento utilizado foi o Cruzamento Uniforme, por gerar a maior combinação entre os participantes do processo, conforme descreve a Seção 5.6.2. O método de mutação utilizado também é semelhante ao utilizado no algoritmo de agrupamento. Primeiramente uma máscara binária é gerada de forma aleatória com o mesmo tamanho do cromossomo do indivíduo a ser mutado. Onde for encontrado um valor igual a 1 na máscara, terá seu valor na posição correspondente no cromossomo modificado para o seu complemento.

6.4 PSEUDOCÓDIGO DA METODOLOGIA

A Figura 6.1 apresenta o pseudocódigo do algoritmo para o cálculo dos pontos atratores.

Devem ser calculados os valores máximos e mínimos de variação dos atributos em cada agrupamento gerado pelo algoritmo apresentado no capítulo 5 para que seja definido o domínio do espaço que contém o agrupamento. O tamanho do cromossomo deve ser calculado e a população inicial gerada. Em seguida, calculados os *fitness* dos indivíduos na população inicial, após esta avaliação, a seleção de pares é realizada e a aplicação dos operadores genéticos.

O Algoritmo Genético deve ser executado durante um determinado número de gerações, e a cada geração os métodos de seleção, cruzamento e mutação devem ser aplicados de acordo com as suas taxas de execução. Este processo deve ser realizado até que o número de gerações chegue ao fim. Ao final da execução do algoritmo o melhor indivíduo encontrado de todas as gerações é obtido com o ponto atrator de densidade calculado onde a densidade em cima deste ponto é máxima.

```
Algoritmo de Calculo de Pontos Atratores
1  Leitura do arquivo de entrada com o agrupamento
2
3  Cálculo dos valores máximos e mínimos da região do agrupamento
4
5  Gerar População inicial
6
7  Para cada indivíduo da População faça
8  |       Iniciar cromossomo aleatoriamente
9  Fim do para
10
11 Para cada indivíduo da População faça
12 |       Converter o cromossomo para valor contínuo
13 |       Avaliar o cromossomo segundo Função (6.2)
14 |       Gerar fitness cromossomo
15 Fim do para
16
17 Para cada geração do algoritmo faça
18 |       Selecionar indivíduos pais na População
29 |       Realizar cruzamento dos pais de acordo com probabilidade
20 |       Gerar nova_geração
21 |       Para cada indivíduo da nova_geração faça
22 |       |       Realizar mutação de acordo com probabilidade
23 |       |       Converter o cromossomo para valor contínuo
24 |       |       Avaliar cromossomo segundo Função (6.2)
25 |       |       Gerar fitness cromossomo
26 |       Fim do para
27 |       Inserir nova_geração da População
28 |       Armazenar Melhor Indivíduo
29 Fim do para
30 Retornar Melhor Indivíduo
```

Figura 6.1. Algoritmo de Calculo de Pontos Atratores em Regiões de Grupos de Dados

7 TESTES E RESULTADOS

Testes foram realizados com o objetivo de estudar a viabilidade do algoritmo de agrupamento e o algoritmo de cálculo de atratores. Os testes incluem o cálculo do agrupamento, onde serão apresentados gráficos com os resultados destes, apresentando visualmente dos pontos pertencentes a cada agrupamento.

Também foram feitos testes do cálculo de pontos atratores da região de cada agrupamento. Serão apresentados gráficos com a posição relativa de cada atrator. Também foram feitos gráficos comparativos da densidade do atrator encontrado com o restante dos pontos pertencentes a base de dados.

7.1 MATERIAIS E MÉTODOS

Para os testes realizados a seguir foram utilizadas duas bases de dados utilizadas na literatura, a base de dados Iris e a base de dados Glass. Estas bases são principalmente utilizadas para testes de algoritmos de agrupamento.

A base de dados Iris contém dados reais multivariados, cujos atributos refletem características medidas sobre as flores de plantas do gênero Iris. Este conjunto contém 150 objetos divididos em 3 grupos, cada uma com 50 objetos e 4 atributos.

A base de dados Glass contém dados reais multivariados sobre tipos de vidro, sendo utilizada para classificar fragmentos de vidro deixados em cenas de crime. Contém 214 objetos com 9 atributos cada, sendo divididos em 6 grupos. A distribuição de objetos por grupos é: 70, 76, 17, 13, 9 e 29 objetos em cada grupo.

Todos os testes foram feitos em um computador Intel® CoreTM2 Duo T5800 2.00GHz, com 2GB de memória RAM, sistema operacional Microsoft Windows XP x64. Foi utilizada a linguagem de programação Java™ e a plataforma Eclipse de desenvolvimento.

Ambas as bases de dados foram divididas em 10 faixas cada atributo, na etapa de configuração. Um número maior de faixas inviabiliza a utilização do método de grade, já que a diminuição no tamanho das células pode fazer com que cada célula contenha exatamente um

objeto, eliminaria algumas das vantagens do algoritmo para agrupamento baseado em densidade e grade que foi proposto. Se o número de faixas for muito pequeno, pode ocorrer de uma célula conter um número relativamente grande de objetos e, consequentemente, será tratada pelo algoritmo como um grande grupo, o que não é o desejado.

Nos algoritmos de agrupamento a população foi composta de 50 indivíduos e foram realizadas 5000 iterações do algoritmo genético, com taxa de cruzamento igual a 0.8 e taxa de mutação igual a 0.1. Na etapa de configuração os atributos foram divididos em 10 faixas cada.

Nos algoritmos de cálculo dos atratores aplicado a base Iris a população foi composta de 50 indivíduos e foram realizadas 5000 iterações do algoritmo genético. A probabilidade de cruzamento é igual a 0.8 e probabilidade de mutação igual a 0.1. O valor utilizado para σ é igual a 0.4. O que quer dizer, que a janela de vizinhança utilizado pela função de densidade e função de influencia foi de 0.4.

Nos algoritmos de calculo dos atratores aplicado a base Glass a população foi composta de 100 indivíduos e foram realizadas 50000 iterações do algoritmo genético. A probabilidade de cruzamento é igual a 0.8 e probabilidade de mutação igual a 0.1. O valor utilizado para σ é igual a 1.2.

7.2 RESULTADOS OBTIDOS DA BASE IRIS

Neste tópico serão discutidos os resultados da utilização do algoritmo genético baseado em densidade e grade aplicado no agrupamento da base de dados Iris.

7.2.1 Resultados Agrupamentos - Iris

As Figuras 7.1-7.3 apresentam os resultados envolvendo os grupos formados pelo algoritmo considerando a base de dados Iris. A Figura 7.4 apresenta os agrupamentos originais das 150 instancias na base considerando os três primeiros atributos.

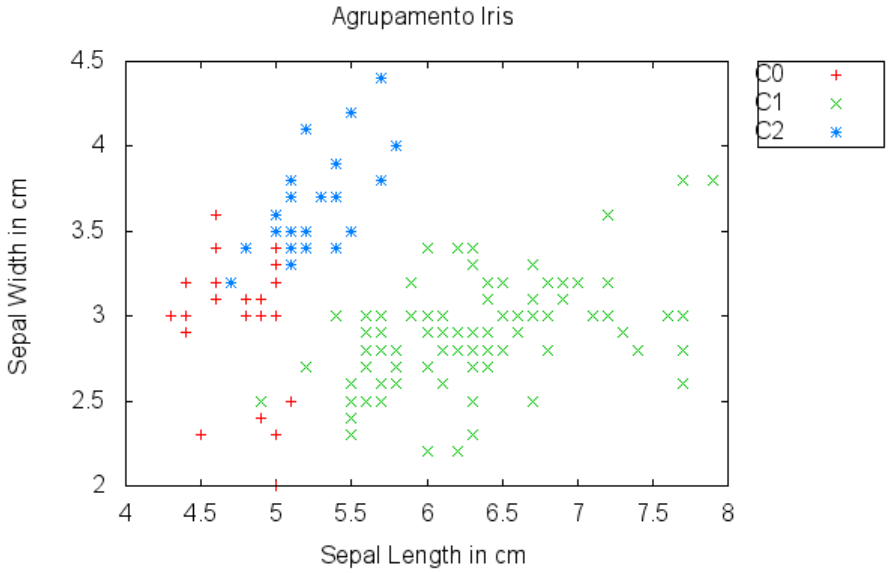

Figura 7.1. Agrupamento gerado para a base de dados Iris (visual utilizando o primeiro e o segundo atributos)

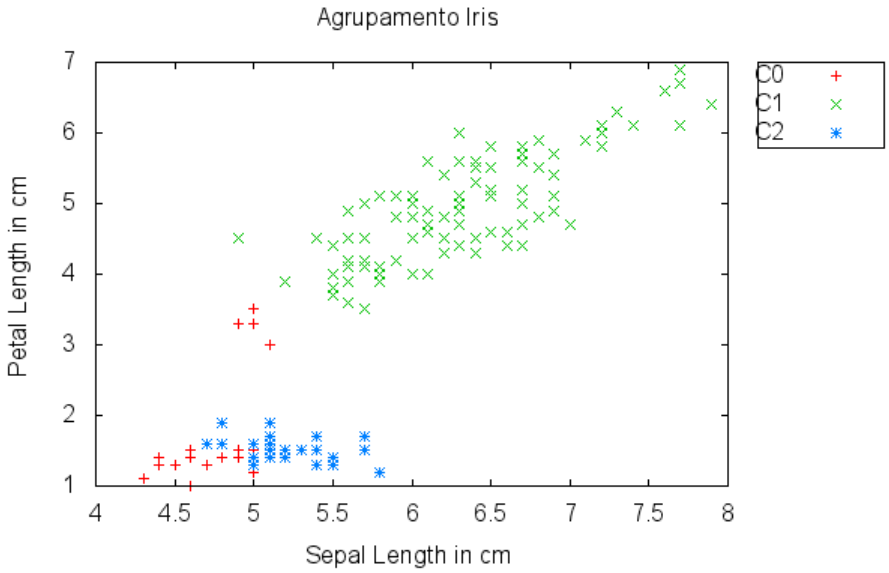

Figura 7.2. Agrupamento gerado para a base de dados Iris (visual utilizando o primeiro e o terceiro atributos)

Observando estes gráficos é possível notar que foram formados três agrupamentos de tamanhos distintos. A estratégia de agrupar por densidade levou à formação de um grande agrupamento por possuir uma grande concentração de pontos. Enquanto que o segundo e terceiro agrupamentos foram formados por pontos que estão mais próximos entre si.

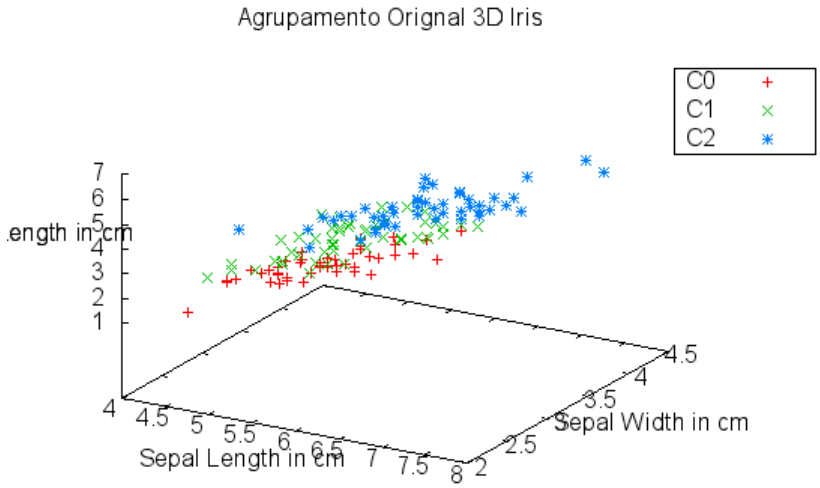

Figura 7.3. Agrupamento 3D da base de dados Iris Original (visual utilizando os 3 primeiros atributos)

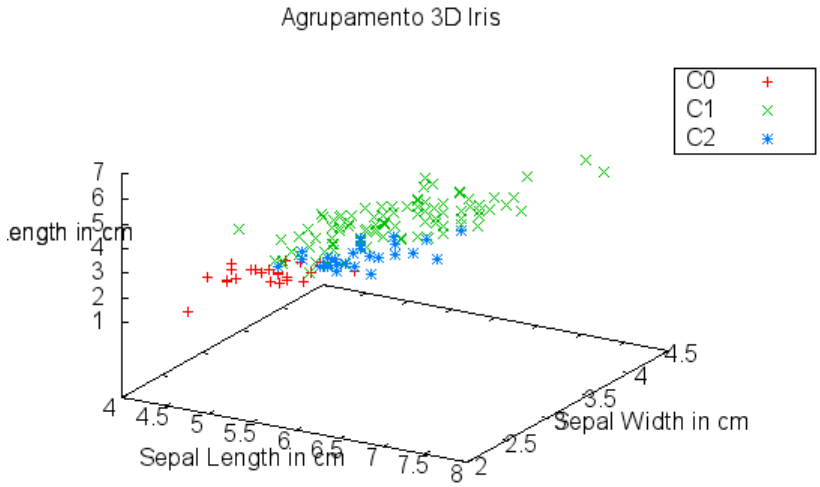

Figura 7.4. Agrupamento 3D da base de dados Iris (visual utilizando os 3 primeiros atributos)

Comparando as Figuras 7.3 e 7.4 percebem-se diferenças entre os agrupamentos originais e formados pelo algoritmo. O agrupamento 1, por exemplo, encontrado pelo algoritmo AGBDG não é exatamente o grupo 1 da base original. A Tabela 7.1 especifica os resultados obtidos e a comparação com o original.

Tabela 7.1 – Comparação dos agrupamentos encontrados pelo método AGBDG com a classificação original da base de dados Iris

Grupo	AGBDG	%	Original	%
1	27	18.0	50	33.3
2	96	64.0	50	33.3
3	27	18.0	50	33.3

Como foi visto, um grande grupo foi formado contendo 96 objetos, e os dois menores possuem 27 objetos cada. Resultado muito diferente do agrupamento original com 50 objetos cada. Os resultados mostram que comparado ao algoritmo original, e provavelmente a outros algoritmos de agrupamento os resultados aparentam não ser eficientes. Mas com relação aos objetivos do trabalho os resultados se mostraram eficientes, pois com o agrupamento por densidade foi formado um agrupamento maior, com 96 objetos, o que quer dizer que estes formam um grande conjunto de pontos similares. Com a minimização das distancias formaram-se os dois grupos menores, com densidades similares. É preferível que o cálculo de pontos atratores ocorra em regiões com alta densidade.

A seguir são apresentados os resultados encontrados no cálculo de atratores na base de dados Iris. Gráficos comparativos de suas funções de influência e densidade de influência foram destacados.

7.2.2 Resultados da Busca de Pontos Atratores - Iris

Os gráficos seguintes apresentam os pontos atratores encontrados na busca realizada na base de dados Iris.

A Tabela 7.2 apresenta os atratores encontrados para os grupos na base de dados Iris.

Tabela 7.2. Atratores encontrados e seus valores respectivos em cada atributo.

	Sepal Length in cm	Sepal Width in cm	Petal Length in cm	Petal Width in cm
A0	**4.81333333**	**3.13548387**	**1.40322581**	**0.16666667**
A1	**5.67419355**	**2.76774194**	**4.09365079**	**1.3**
A2	**5.14**	**3.52**	**1.5**	**0.24285714**

7.2.2.1 Posição Relativa Dos Atratores No Espaço

Os gráficos seguintes apresentam as posições espaciais relativas a estes atratores para os dois primeiros atributos. Em cada gráfico é utilizado somente o grupo a que o atrator pertence.

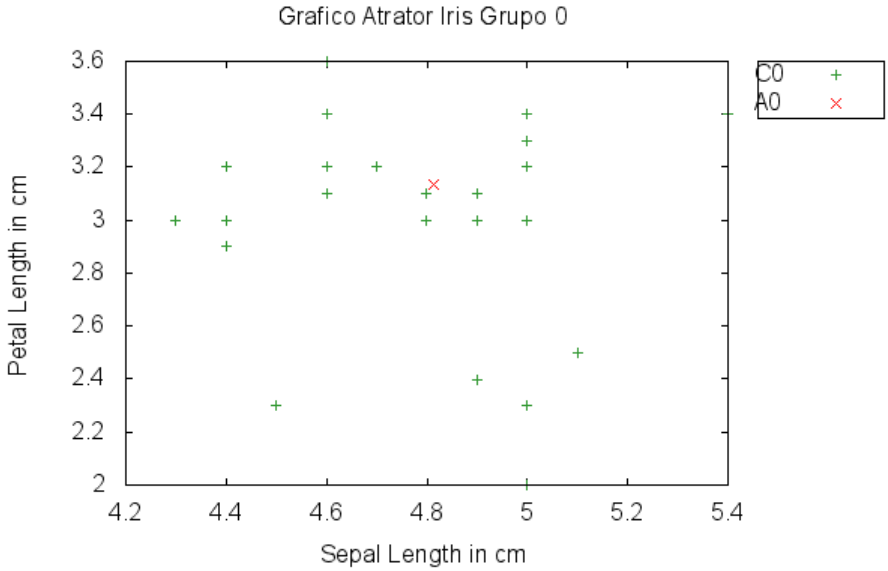

Figura 7.5. Gráfico do atrator A0 encontrado, e sua posição em seu respectivo grupo C0.

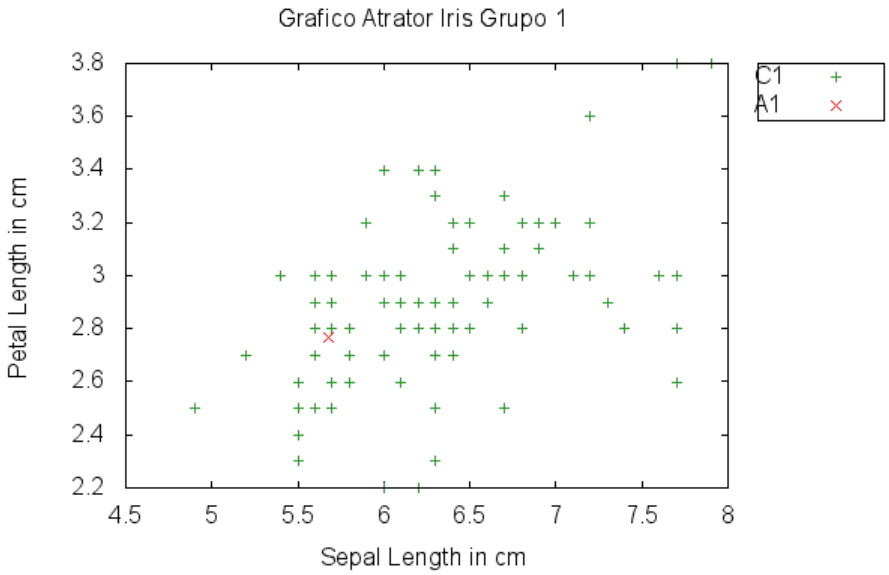

Figura 7.6. Gráfico do atrator A1 encontrado, e sua posição em seu respectivo grupo C1.

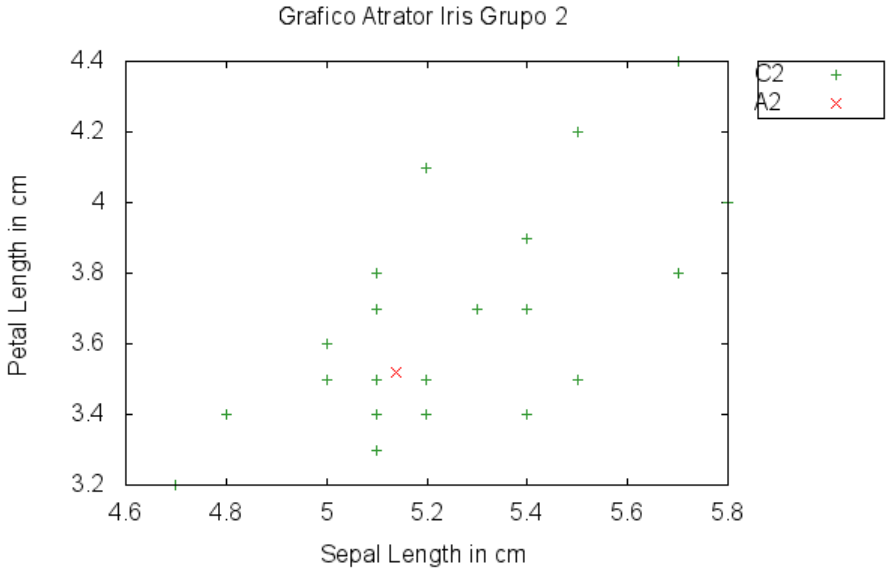

Figura 7.7. Gráfico do atrator A2 encontrado, e sua posição em seu respectivo grupo C2.

Como é possível notar, os atratores encontrados se posicionam próximos aonde há a maior concentração de pontos em seus respectivos grupos, respeitando a janela de 0,4 para a vizinhança. Em cada execução do algoritmo de cálculo dos atratores é encontrado somente 1 atrator em cada grupo, o que não impede de, caso seja realizado uma nova busca, no agrupamento maior, seja encontrado um segundo ponto atrator. Os resultados estão dentro do definido na teoria.

A seguir os gráficos de densidade dos atratores serão apresentados, comparando as suas densidades com os pontos da base de dados como um todo, não somente os pertencentes aos grupos em que foram calculados.

7.2.2.2 Gráficos de Densidade da Base de Dados em Comparação com os Atratores

Nesta seção serão apresentados alguns gráficos relativos à comparação da densidade calculada dos pontos da base de dados com os atratores encontrados.

Nestes gráficos os atratores foram posicionados na origem do gráfico, e a partir daí, os outros pontos da base foram posicionados utilizando a distancia destes em relação ao atrator encontrado. A mesma função de densidade de influência utilizada para o calculo do atrator é utilizada para calcular a densidade dos outros pontos. Este gráfico é utilizado para mostrar a qualidade da densidade do atrator encontrado, em comparação aos outros pontos da base.

Nas figuras 7.8 a 7.10 nota-se que os atratores encontrados não possuem as maiores densidades da base, em comparação aos outros dados.

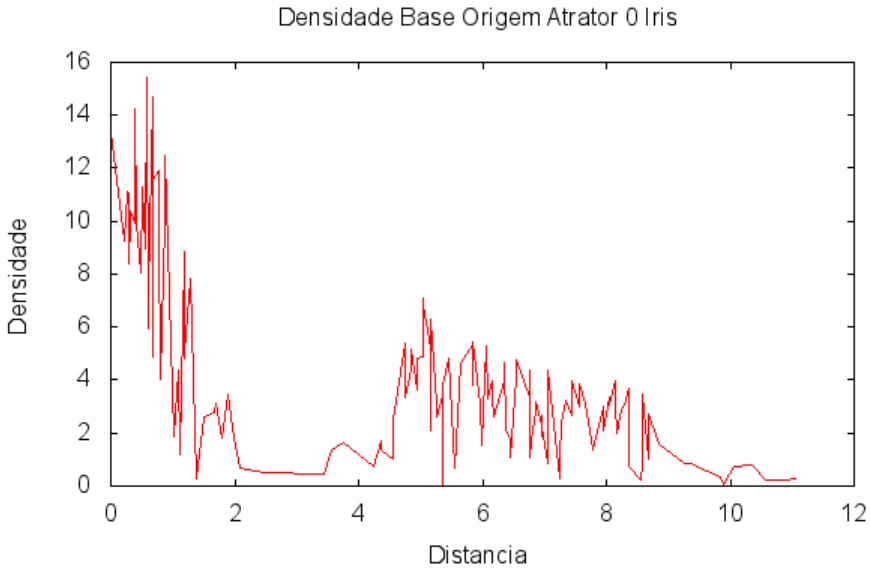

Figura 7.8. Gráfico comparativo de densidades dos pontos da base de dados com o atrator A0 encontrado.

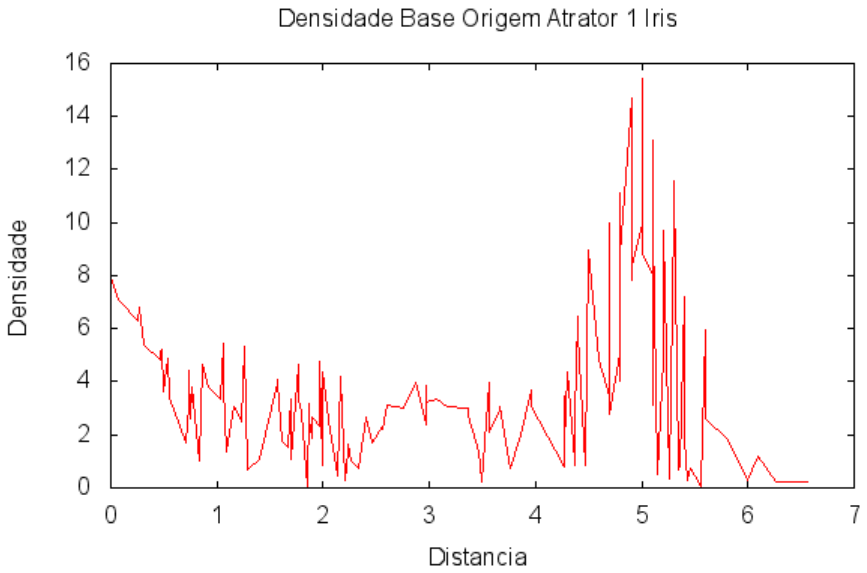

Figura 7.9. Gráfico comparativo de densidades dos pontos da base de dados com o atrator A1 encontrado.

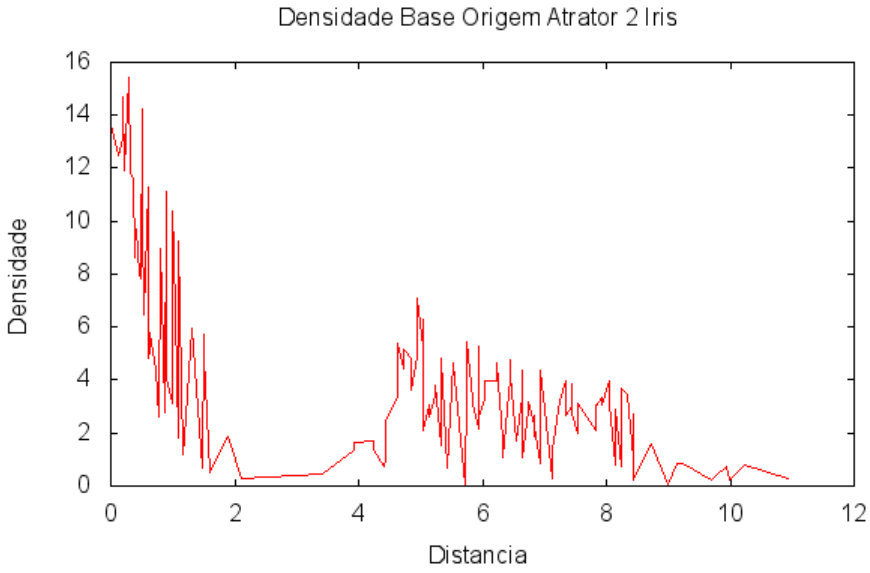

Figura 7.10. Gráfico comparativo de densidades dos pontos da base de dados com o atrator A2 encontrado.

Os gráficos anteriores mostram que os atratores encontrados não apresentam os maiores valores de função densidade em comparação aos pontos na base como um todo. Como é possível notar dois destes atratores se encontram próximos às regiões de alta densidade. Isto era esperado já que os atratores configuram-se em máximos locais da função densidade. A seção seguinte apresenta os gráficos de densidade dos atratores sobre seus respectivos grupos.

7.2.2.3 Gráficos de Densidade dos Grupos encontrados em Comparação com os seus Atratores Respectivos

Os gráficos seguintes apresentam um comparativo das densidades dos pontos em cada grupo, tendo como ponto de origem o ponto atrator correspondente. Mais especificamente, a densidade é calculada utilizando todos os pontos do grupo do atrator. Os resultados são comparados com a densidade do atrator encontrado, e dispostos utilizando a distância em relação ao mesmo. Neste caso, nota-se que o atrator encontrado possui a maior densidade em comparação com os outros pontos.

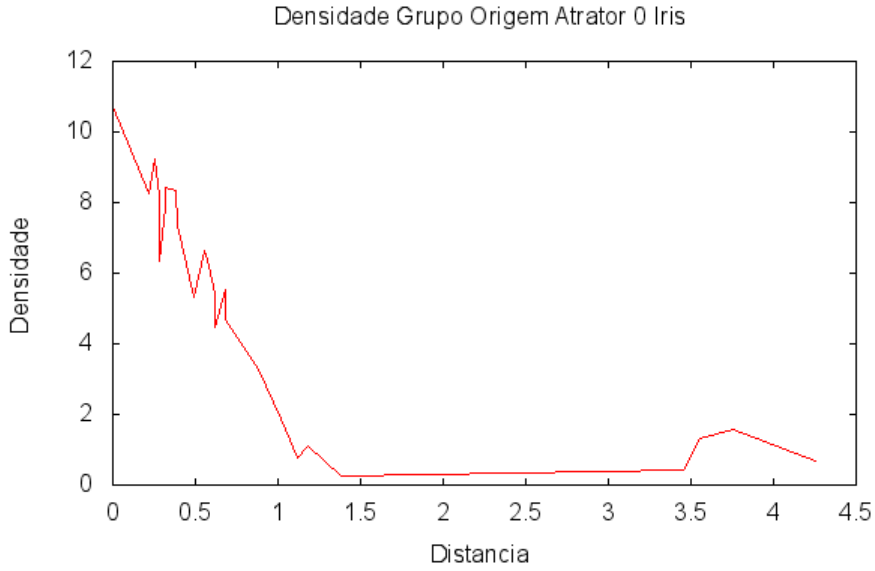

Figura 7.11. Gráfico comparativo de densidades dos pontos do agrupamento C0 de dados com o atrator A0 encontrado.

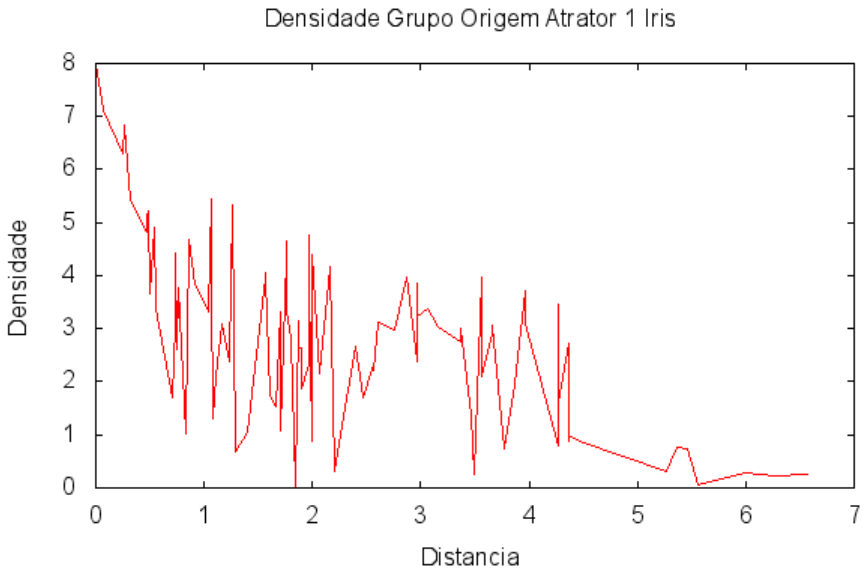

Figura 7.12. Gráfico comparativo de densidades dos pontos do agrupamento C1 de dados com o atrator A1 encontrado.

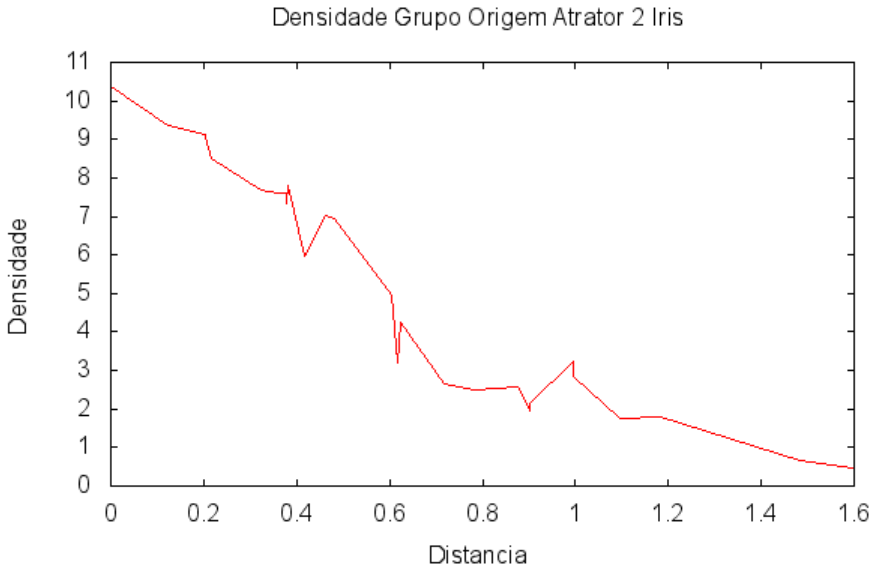

Figura 7.13 Gráfico comparativo de densidades dos pontos do agrupamento C2 de dados com o atrator A2 encontrado

Como foi possível verificar os resultados das densidades dos atratores encontrados foram os maiores de seus respectivos grupos. As regiões mais próximas dos atratores possuem alta densidade. Isto leva a crer que nestes agrupamentos há uma tendência aos pontos estarem próximos aos atratores. Esta figura equivale ao segundo agrupamento, o que possui a maior densidade, mesmo a certa distância do atrator encontrado há outros pontos de densidade de influência alta. Caso seja realizada uma nova busca de atratores nessa região, possivelmente poderão ser encontrados outros atratores.

Nos agrupamentos C1 e C2 por possuírem densidades menores, então a busca por atratores se torna menos complicada. Possivelmente, caso seja realizada uma nova busca, a posição do atrator não mudará, ou será bem próxima a atual.

Mesmo pontos próximos aos atratores encontrados possuem alta densidade de influência, pois estes estão em uma região onde a concentração de dados é grande, por isso as suas densidades de influencia são próximas aos atratores.

A seguir serão apresentados os gráficos de influencia dos atratores.

7.2.2.4 Gráficos de Influencia do Atrator em Relação a Base de Dados

Estes gráficos apresentam a influencia do atrator sobre a base de dados Iris, tendo como ponto de origem o atrator, nota-se que os pontos mais próximos sofrem maior influencia. Uma importante notação é que a influencia de um ponto é maior caso a distancia estimada esteja dentro da região determinada por σ. Com isso a influencia sobre um ponto é mais alta, conforme este se aproxima do atrator aferido. Para o calculo desta influência foi utilizada a Função (3.5), que nos permite calcular a influência de um ponto sobre outro.

A influencia do atrator é calculada utilizando todos os pontos da base de dados. Com isto é possível verificar a atenuação da influência do atrator a medida que os pontos da base se distanciam.

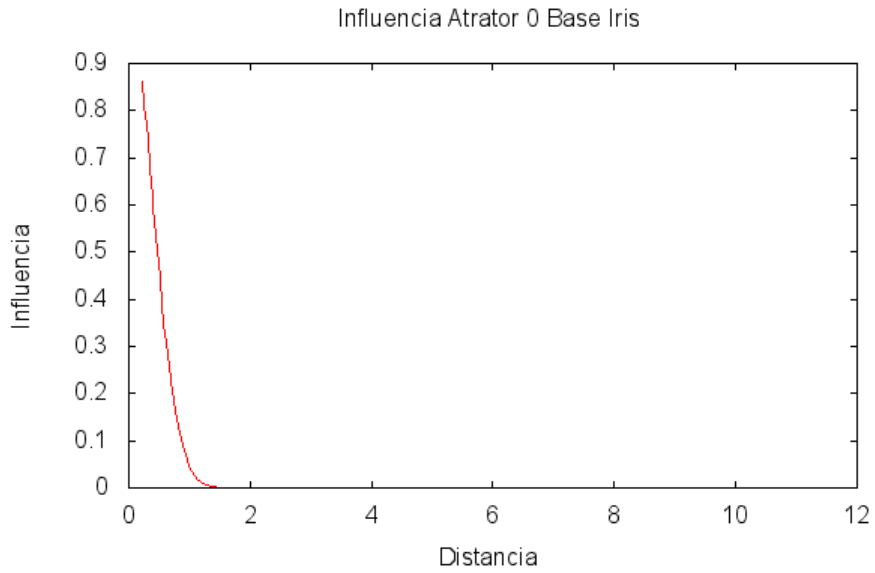

Figura 7.14. Gráfico comparativo de influencia do atrator A0 sobre a base de dados

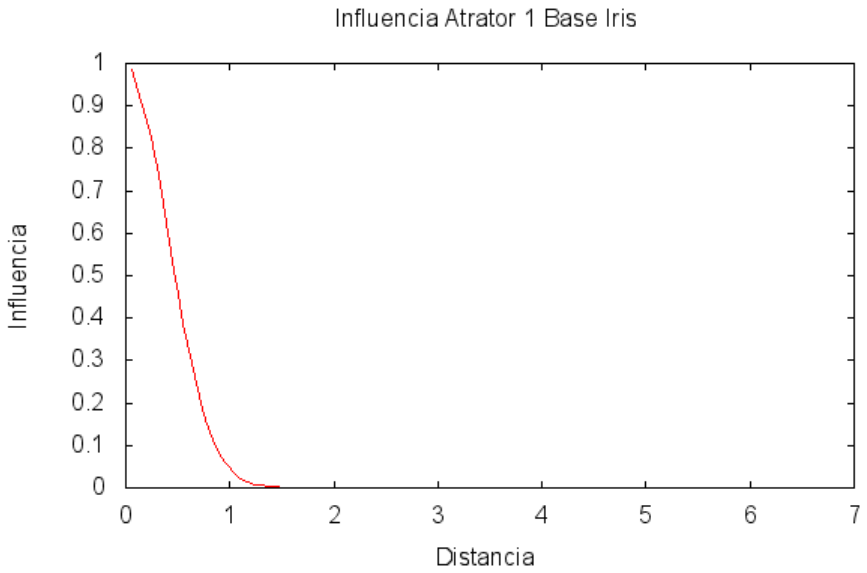

Figura 7.15. Gráfico comparativo de influencia do atrator A1 sobre a base de dados

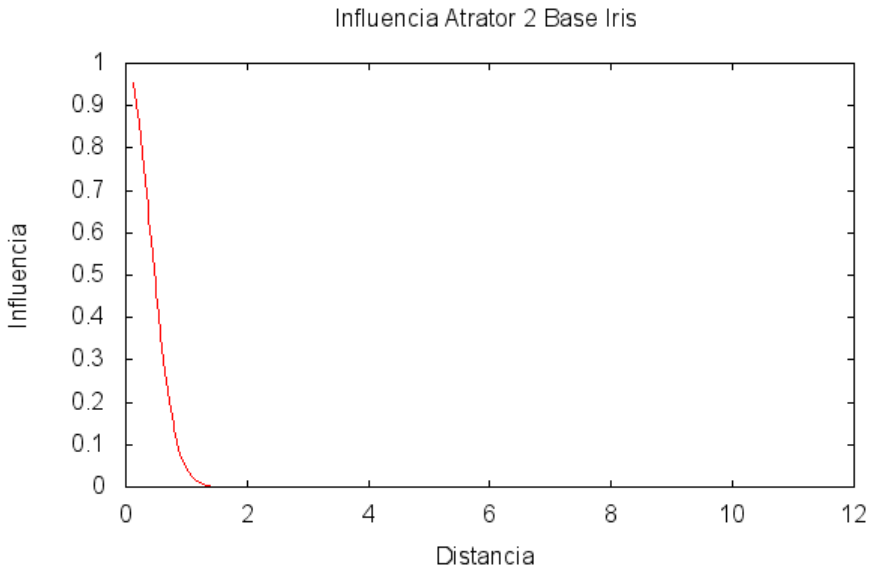

Figura 7.16. Gráfico comparativo de influencia do atrator A2 sobre a base de dados

Os gráficos anteriores mostram a influência exercida pelo atrator sobre os pontos da base de dados. À medida que a distância dos pontos para o atrator aumenta a influência do atrator sobre estes diminui. Com este gráfico é possível estudar a atenuação da influência do atrator sobre a base de dados.

Á seguir os gráficos de influencia do atrator em relação a seu grupo.

7.2.2.5 Gráfico de Influencia do Atrator Sobre o Grupo a que Pertence

Os gráficos seguintes apresentam as influencias do atrator sobre os grupos a que pertencem. Foi utilizada a mesma métrica de distancia para a representação dos pontos em estudo. Com isto é possível verificara atenuação da influência do atrator sobre o agrupamento a que pertence a medida que os pontos se distanciam do atrator.

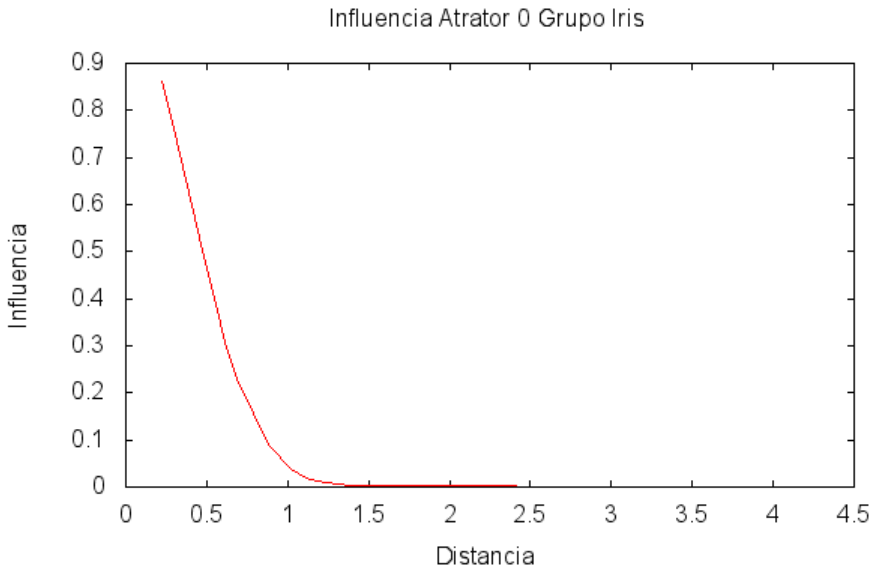

Figura 7.17. Gráfico comparativo de influencia do atrator A0 sobre o agrupamento C0 encontrado

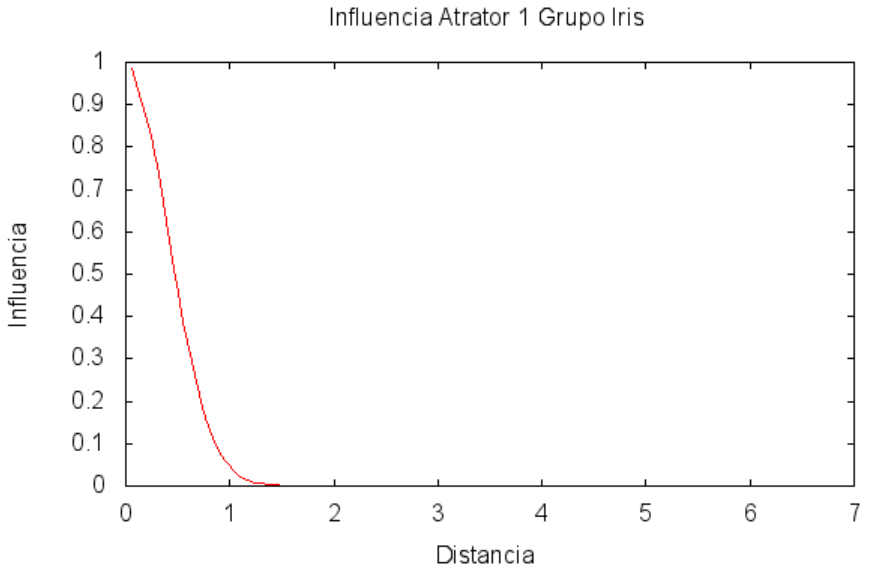

Figura 7.18. Gráfico comparativo de influencia do atrator A1 sobre o agrupamento C1 encontrado

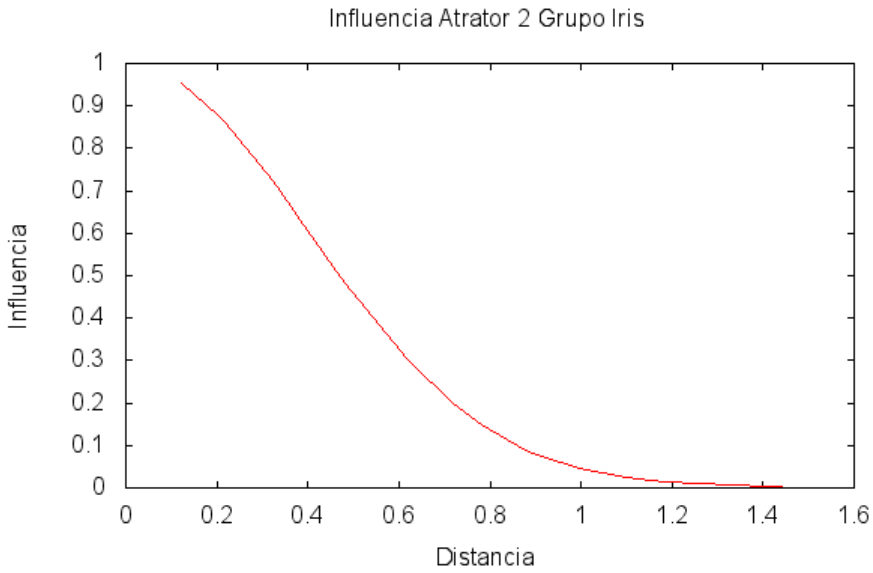

Figura 7.19. Gráfico comparativo de influencia do atrator A1 sobre o agrupamento C1 encontrado

Os resultados apresentados da influência dos atratores sobre os pontos do grupo no qual foi calculado estão dentro do esperado, assim como foi visto nos gráficos da influência do atrator sobre a base de dados, a influência decai a medida que os pontos se afastam do atrator.

Tabela 7.3. Acumulação de pontos na região de influência em unidades de distância.

	Número de Pontos		
Segmentos de Área de Influência	A0	A1	A2
Até 1.00	19	19	23
Até 3.00	23	67	27
Até 5.00	27	89	
Até 7.00		96	

A Tabela 7.3 apresenta a acumulação de pontos, a medida que a distancia para o atrator aumenta. Nota-se que o número de pontos com distância menos do que 1.00 é alta. A conclusão que é possível tirar disso é que o atrator se posiciona em regiões de grande concentração de pontos. Esta tabela apresenta o poder de atratividade do atrator em relação ao seu grupo, sua capacidade de atrair diversos pontos para a sua região de influência.

Os atratores A0 e A2 conseguiram acumular a maior número de pontos em uma área menos do que 1.00 unidades de distancia. O atrator A1 acumulou seu maior percentual a até 3.00, mas também possui uma grande quantidade em uma distância menor do que 1.00.

7.3 RESULTADOS OBTIDOS DA BASE GLASS

Neste tópico serão discutidos os resultados encontrados na busca realizada na base de dados Glass.

7.3.1 Resultados Agrupamentos – Glass

Os resultados seguintes apresentam uma discussão sobre os grupos encontrados na base de dados Glass. Comparações serão feitas com relação a classificação original presente na base.

No agrupamento gerado foram encontrados 6 grupos com 106, 2, 18, 4, 22 e 62 dados em cada, respectivamente

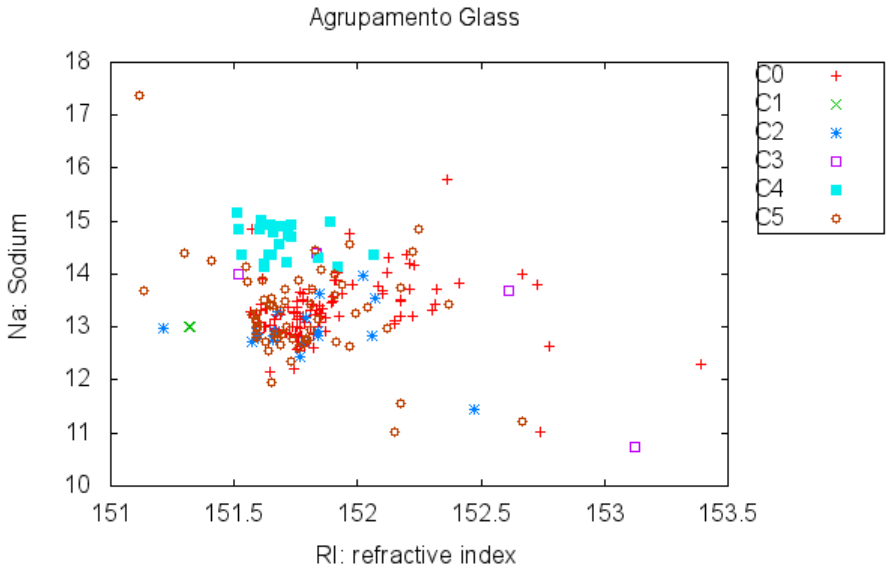

Figura 7.20. Agrupamento gerado para a base de dados Glass (visual utilizando o primeiro e o segundo atributos)

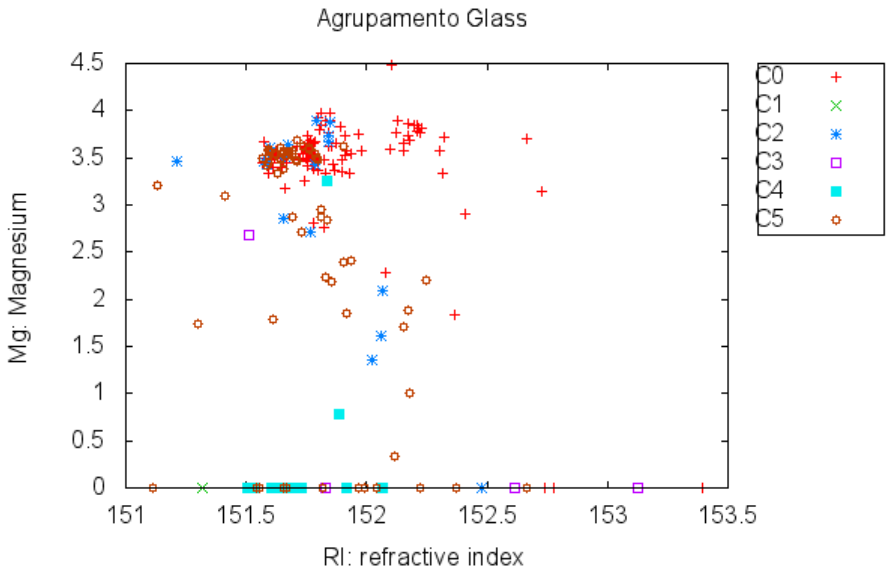

Figura 7.21. Agrupamento gerado para a base de dados Glass (visual utilizando o primeiro e o terceiro atributos)

Utilizando estes gráficos é possível notar que os pontos da base de dado Glass, por possuírem uma grande quantidade de atributos, estão muito próximos um dos outros quando utilizado somente 2 destes atributos para uma apresentação visual. Apesar de ainda assim ser possível notar uma grande concentração de dados não é possível dizer se nos outros atributos o mesmo ocorre.

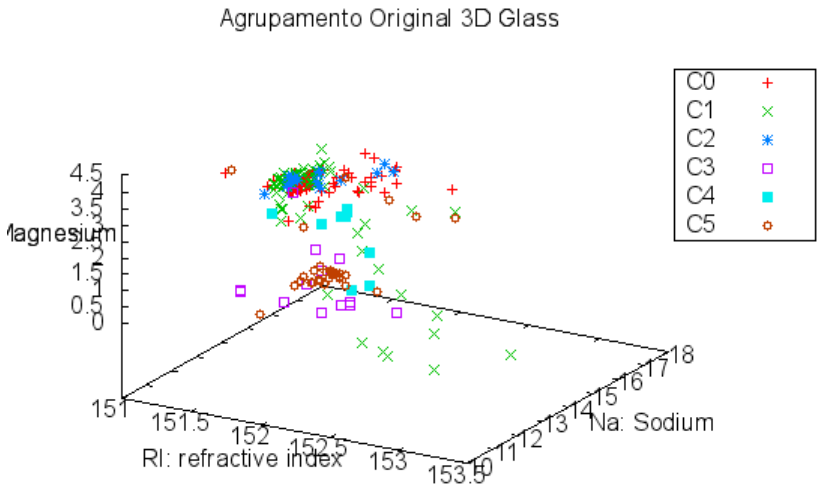

Figura 7.22. Agrupamento 3D da base de dados Glass Original (visual utilizando os 3 primeiros atributos)

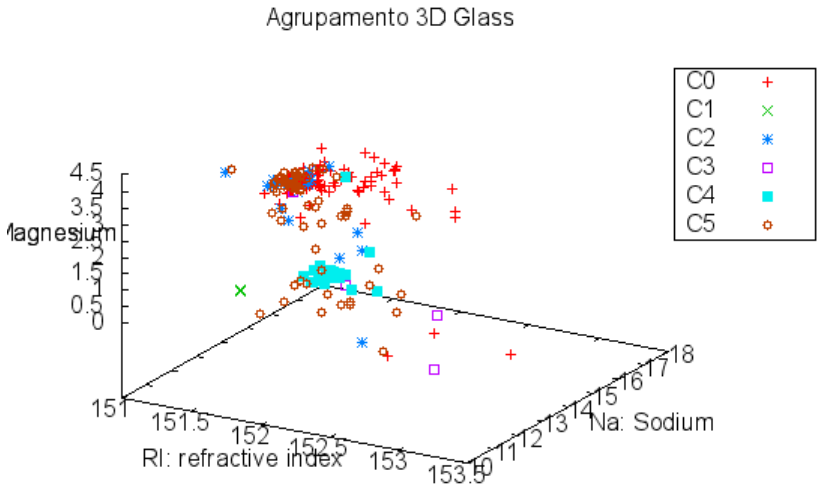

Figura 7.23. Agrupamento 3D da base de dados Glass (visual utilizando os 3 primeiros atributos)

Nas figuras 7.22 e 7.23 é apresentado visualmente a representação dos três primeiros atributos em terceira dimensão. Já é possível analizar que o agrupamento gerado pelo AGBDG em certos locais da base conseguiu agrupar os dados de forma semelhante, apesar de ainda não serem completamente iguais.

Tabela 7.4 – Comparação dos agrupamentos encontrados pelo método AGBDG com a classificação original da base de dados Glass.

Grupos	AGBDG	%	Original	%
0	106	49.5	70	32.7
1	2	0.9	76	35.5
2	18	8.4	17	7.9
3	4	1.9	13	6.1
4	22	10.3	9	4.2
5	62	29.0	29	13.6

Como é mostrado na tabela todos os agrupamentos foram diferentes dos grupos originais. Estes resultados podem ser interpretados da seguinte forma: pode haver uma grande proximidade de pontos em determinados atributos, o que forma o grupo maior. Enquanto que há grupos menores, possivelmente mais separados dos demais.

Visualmente é possível identificar a grande densidade na região do primeiro agrupamento. Como a busca é agrupar por densidade esta é considerada satisfatória. Mesmo pontos mais espaçados estão no mesmo agrupamento, o que pode ser justificado pelo valor dos demais atributos. O agrupamento 5 (cinco) também é posicionado em uma região de alta densidade, o que justifica a sua densidade. Quando aos demais agrupamentos há uma tentativa de agrupa-los por densidade com o mínimo possível de distancia entre os pontos.

O tópico a seguir apresenta os resultados encontrados na busca por atratores na base de dados Iris. Gráficos comparativos de suas densidades e de influencias serão apresentados.

7.3.2 Resultados da Busca de Pontos Atratores – Glass

Os resultados a seguir apresentam os pontos atratores encontrados utilizando o algoritmo de calculo de pontos atratores.

Ao todo foram encontrados 6 pontos atratores, um para cada agrupamento realizado. Estes estão listados na tabela a seguir.

Tabela 7.5. Atratores encontrados e seus valores respectivos em cada atributo na base Iris.

	Refractive Index	Sodium	Magnesium	Aluminum	Silicon	Potassium	Calcium	Barium	Iron
A0	151.78822521	13.01761252	3.6464775	1.35647059	72.66714286	0.60503937	8.36125122	0.01976471	0.12333333
A1	151.31790476	13.0	0.0	3.02666667	70.57225806	6.21	6.94	0.0	0.0
A2	151.61185039	12.93823529	3.52857143	1.42716535	73.09898039	0.67	8.29482893	0.01741935	0.29412698
A3	151.62103273	13.83849315	2.64853229	3.46643137	69.92937378	1.44627451	5.89959922	2.42876712	0.09032258
A4	151.74049169	14.83984252	0.01913894	1.97338583	73.25505882	0.01717647	8.59896282	1.62282353	0.03
A5	151.48276453	13.21494624	3.51436399	1.55129412	73.01371457	0.43548387	8.52237537	0.0	0.14451613

7.3.2.1 Posição Relativa Dos Atratores No Espaço

Os gráficos seguintes apresentam as posições espaciais relativas a estes atratores para alguns atributos.

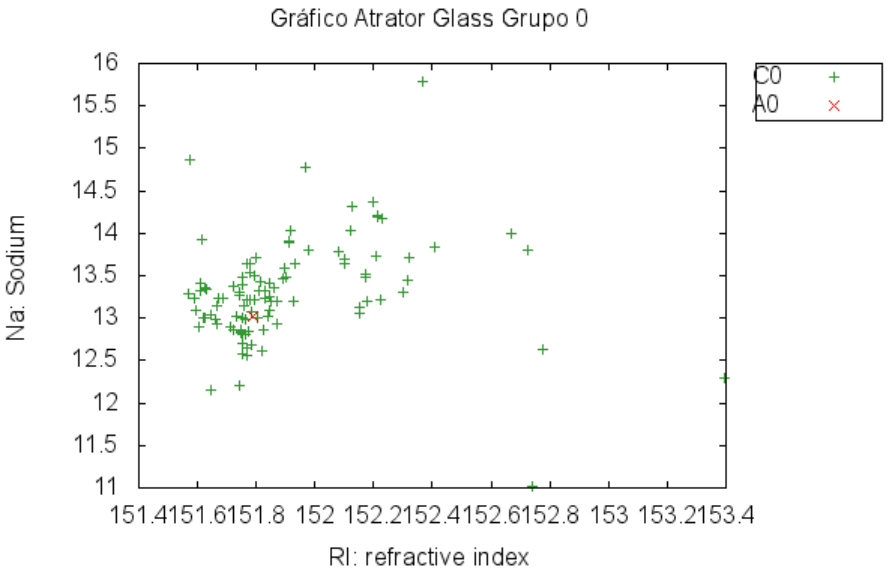

Figura 7.24. Gráfico do atrator A0 encontrado, e sua posição em seu respectivo grupo C0.

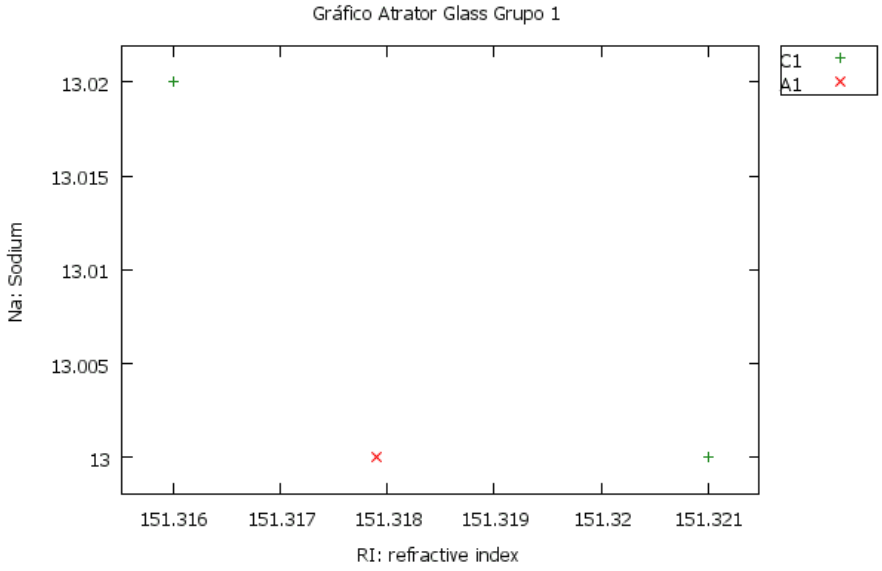

Figura 7.25. Gráfico do atrator A1 encontrado, e sua posição em seu respectivo grupo C1.

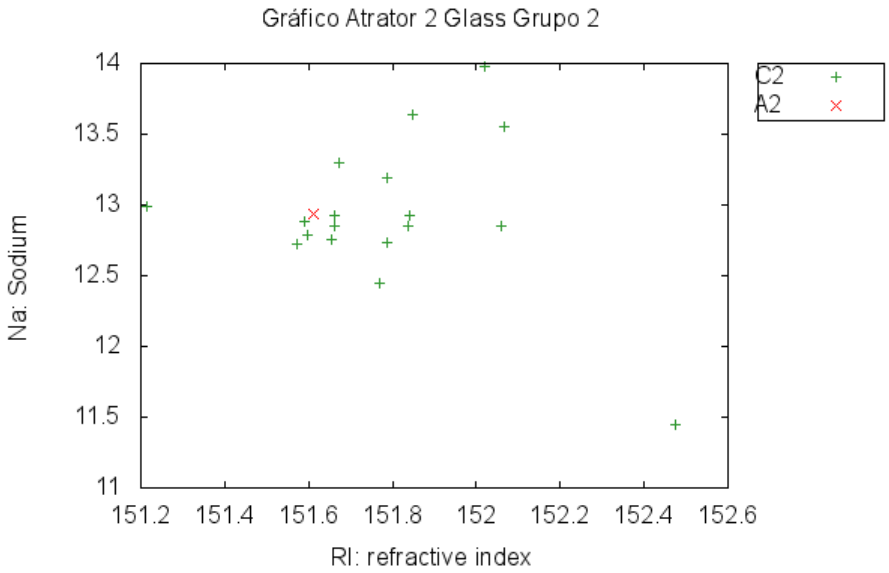

Figura 7.26. Gráfico do atrator A2 encontrado, e sua posição em seu respectivo grupo C2.

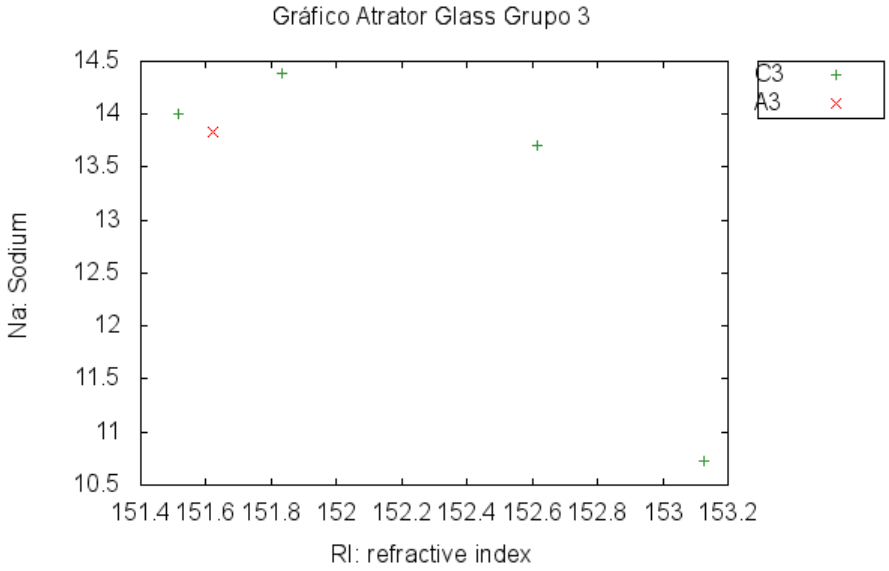

Figura 7.27. Gráfico do atrator A3 encontrado, e sua posição em seu respectivo grupo C3.

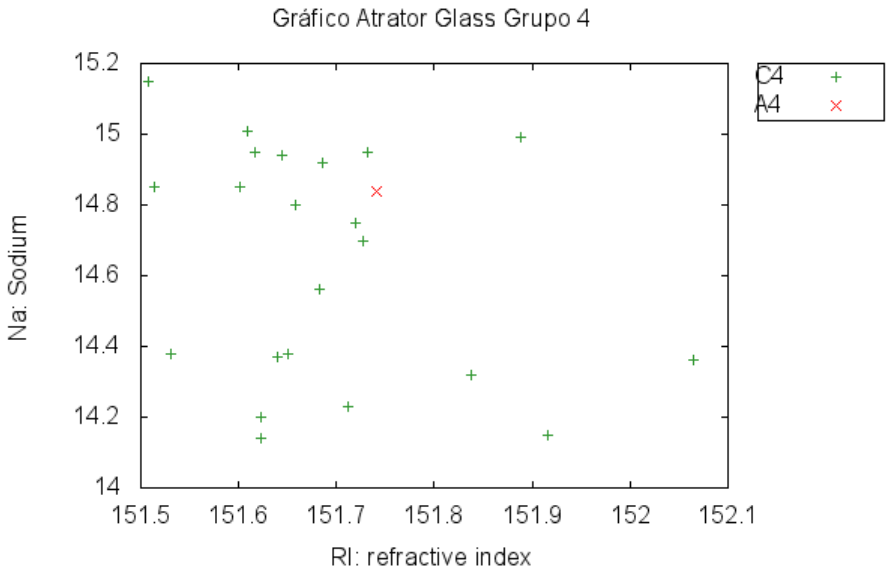

Figura 7.28. Gráfico do atrator A4 encontrado, e sua posição em seu respectivo grupo C4.

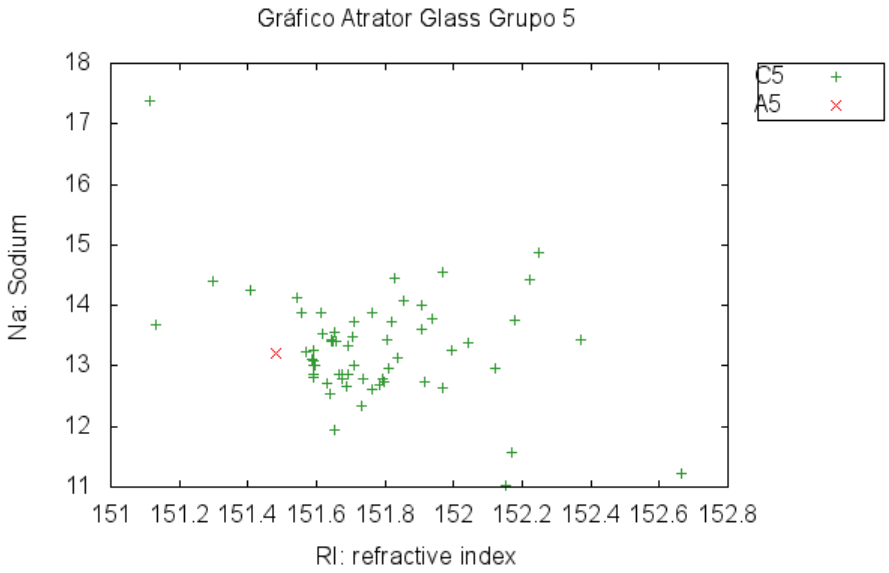

Figura 7.29. Gráfico do atrator A5 encontrado, e sua posição em seu respectivo grupo C5.

Os atratores encontrados também estão posicionados próximos a espaços de alta densidade em seus respectivos grupos, respeitando a janela de vizinhança de 0,4. Para os atratores A4 e A5 nas Figuras 7.28 e 7.29 não estarem no centro do espaço de maior densidade significa que os atributos seguintes aos utilizados nos gráficos modificam a sua posição.

O tópico seguinte apresenta os gráficos de densidade dos atratores em comparação com a base de dados Glass. Será possível notar que os seus resultados foram bem diferentes dos resultados aferidos com a base de dados Iris.

7.3.2.2 Gráficos de Densidade de Influência da Base de Dados em Comparação com os Atratores

Os gráficos a seguir apresentam um comparativo das densidades dos pontos da base de dados Glass com os atratores encontrados. Tendo como pontos de origem os atratores e suas densidades de influência, é possível notar que os atratores possuem densidades de influência boas em comparação aos outros pontos de dados apesar de existirem outros pontos com densidades de influência maiores.

A densidade é calculada utilizando todos os dados da base de dados. Os resultados são comparados com a densidade do atrator encontrado, e dispostos utilizando a distancia em relação ao mesmo.

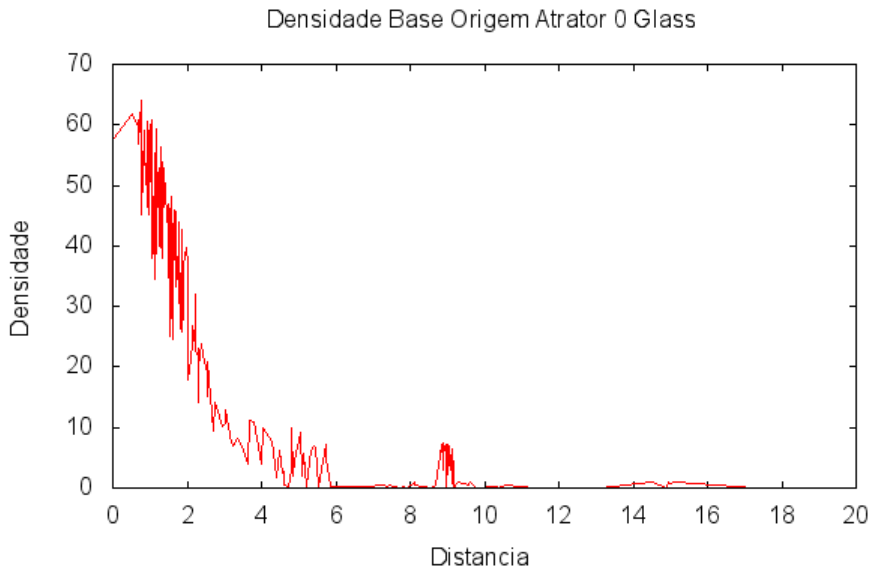

Figura 7.30. Gráfico comparativo da densidade de influência da base de dados em relação a densidade do atrator A0 encontrado

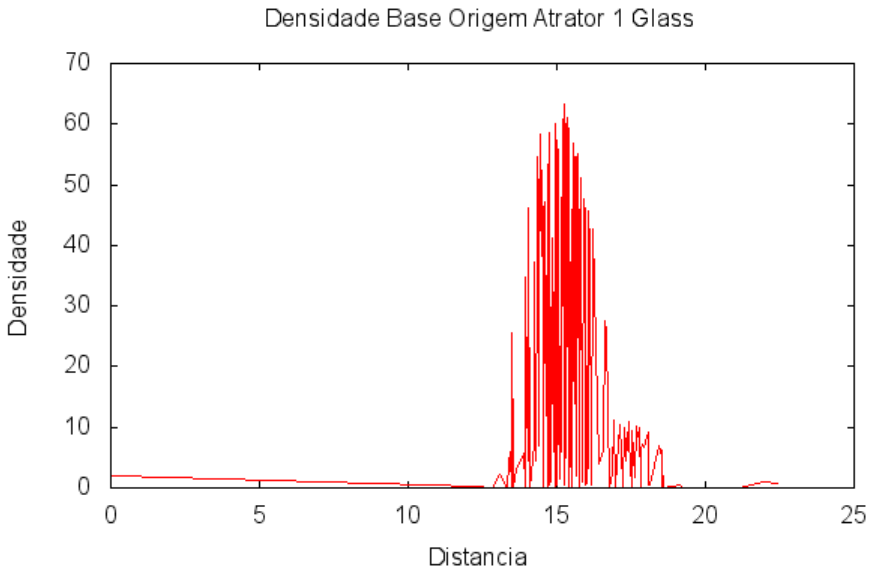

Figura 7.31. Gráfico comparativo da densidade de influência da base de dados em relação a densidade do atrator A1 encontrado

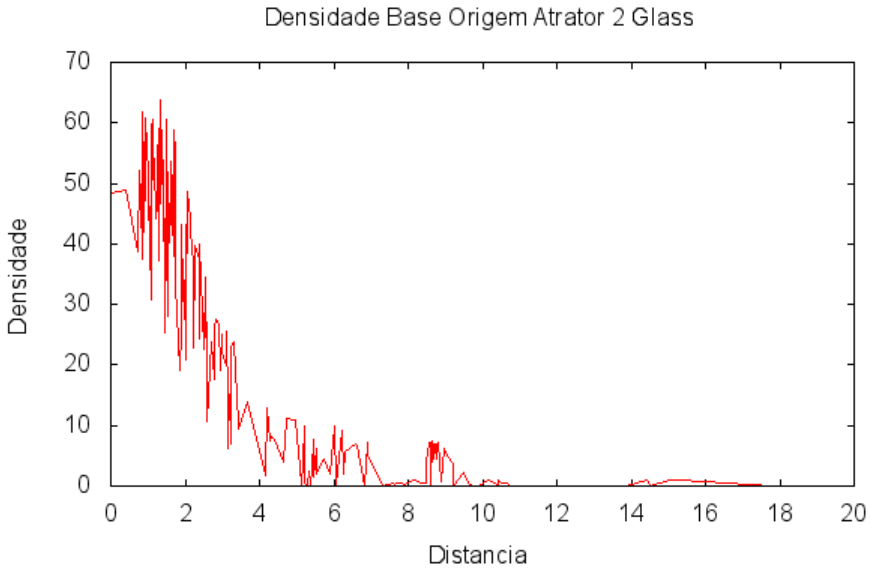

Figura 7.32. Gráfico comparativo da densidade de influência da base de dados em relação a densidade do atrator A2 encontrado

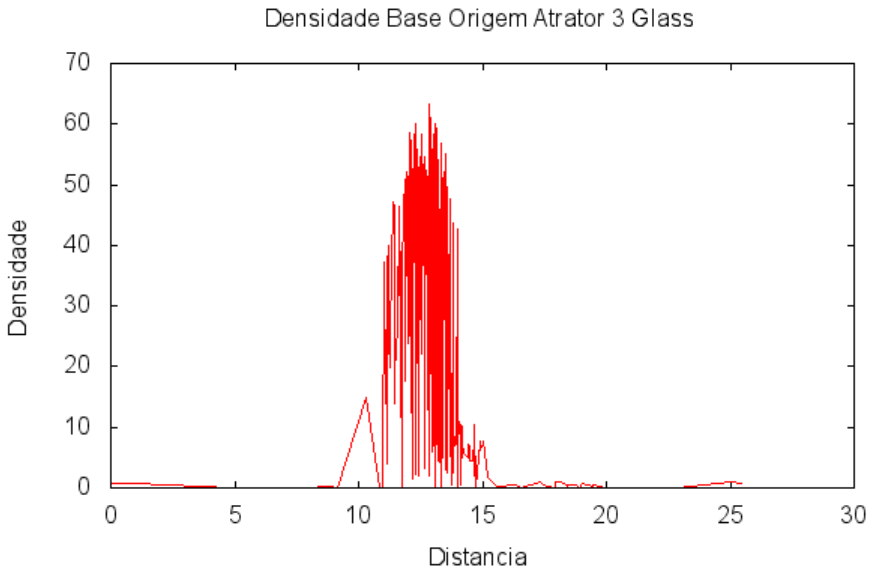

Figura 7.33. Gráfico comparativo da densidade de influência da base de dados em relação a densidade do atrator A3 encontrado

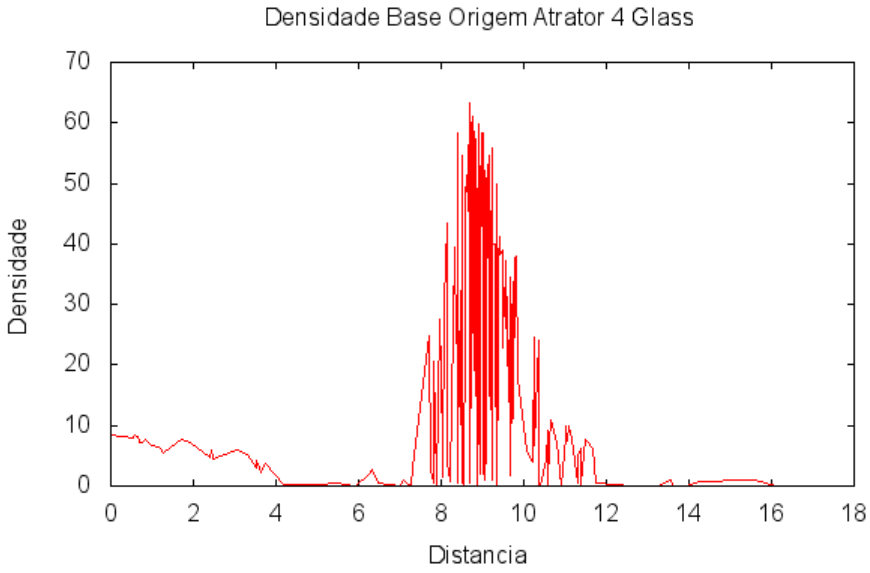

Figura 7.34. Gráfico comparativo da densidade de influência da base de dados em relação a densidade do atrator A4 encontrado

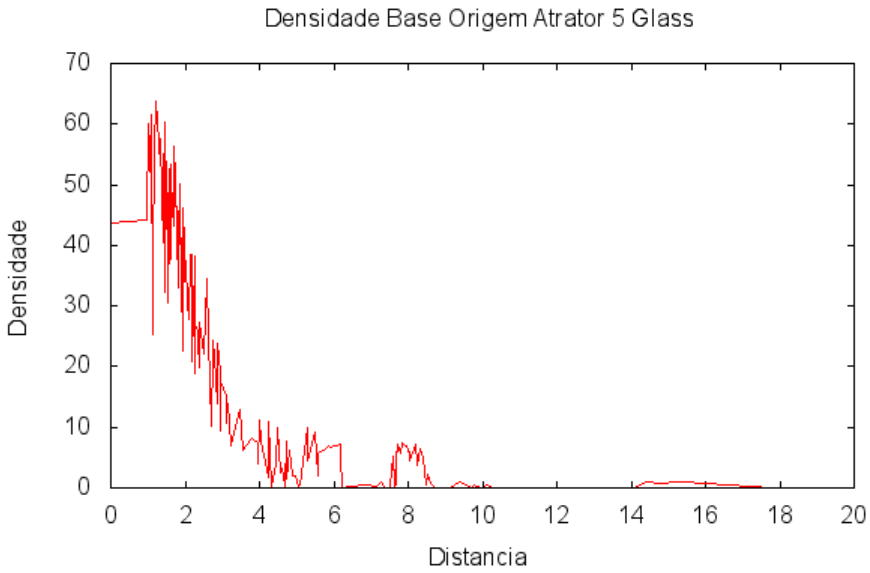

Figura 7.35. Gráfico comparativo da densidade de influência da base de dados em relação a densidade do atrator A5 encontrado

Os pontos atratores encontrados, novamente, não foram os maiores da base de dados. Mas com estes gráficos foi possível notar uma área de pontos com grande densidade. O atrator A0 encontrou um ponto próximo a esta região. O atrator A5 também está próximo a esta região de alta densidade, os demais atratores, por estarem localizados em regiões de baixa densidade, apresentaram resultados menores.

A seguir o tópico relativo a densidade de cada atrator em relação ao grupo a que foi calculado.

7.3.2.3 Gráficos de Densidade de Influência dos Grupos encontrados em Comparação com os seus Respectivos Atratores

A densidade é calculada utilizando todos os dados do grupo do atrator. Os resultados são comparados com a densidade do atrator encontrado, e dispostos utilizando a distancia em relação ao mesmo.

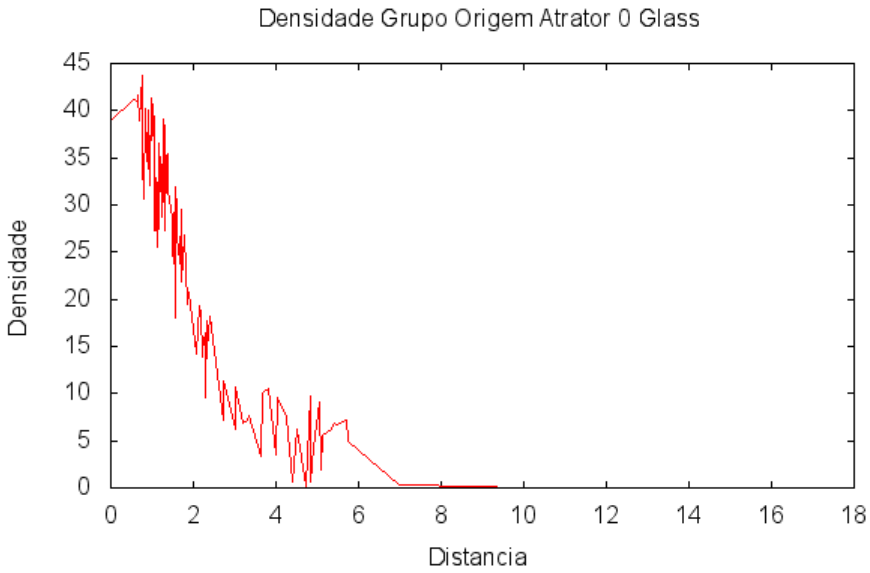

Figura 7.36. Gráfico comparativo da densidade de influência do agrupamento C0 em relação a densidade de influência do atrator A0 encontrado

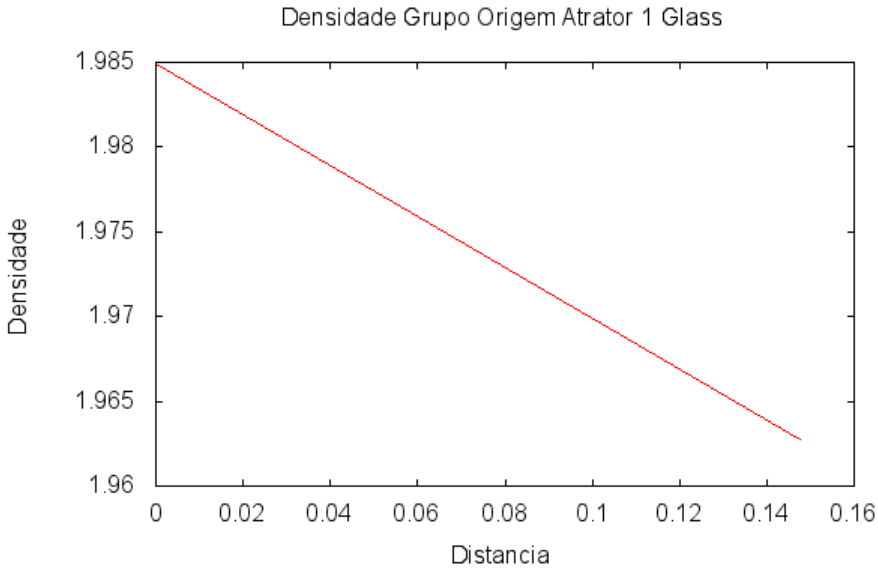

Figura 7.37. Gráfico comparativo da densidade de influência do agrupamento C1 em relação a densidade de influência do atrator A1 encontrado

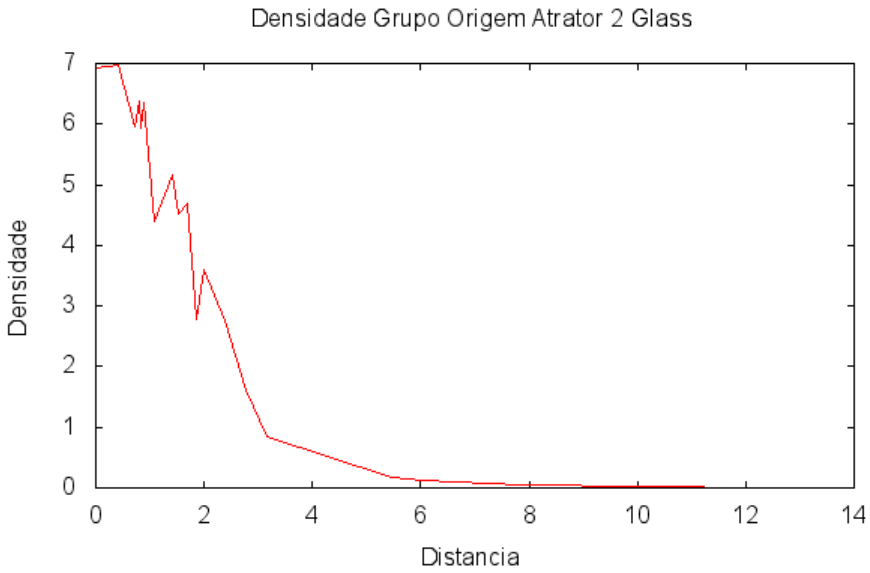

Figura 7.38. Gráfico comparativo da densidade de influência do agrupamento C2 em relação a densidade de influência do atrator A2 encontrado

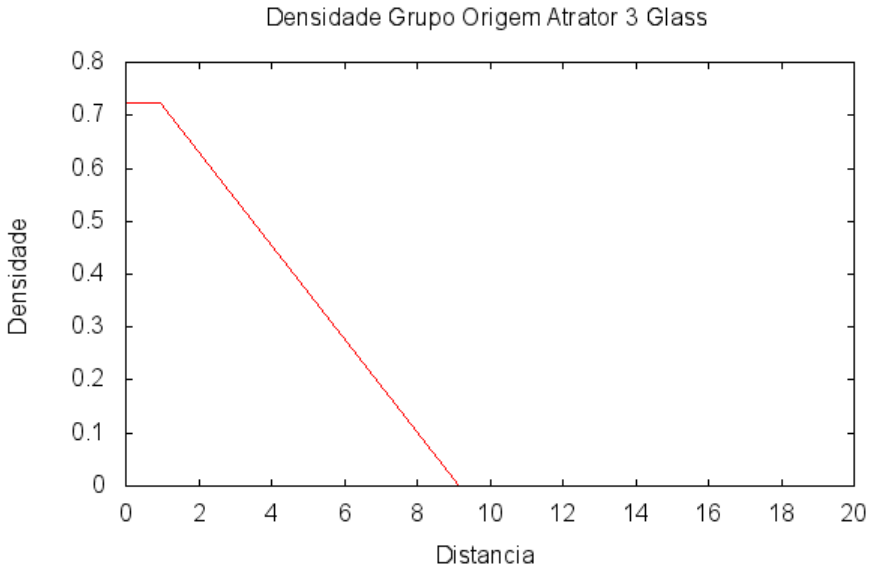

Figura 7.39. Gráfico comparativo da densidade de influência do agrupamento C3 em relação a densidade de influência do atrator A3 encontrado

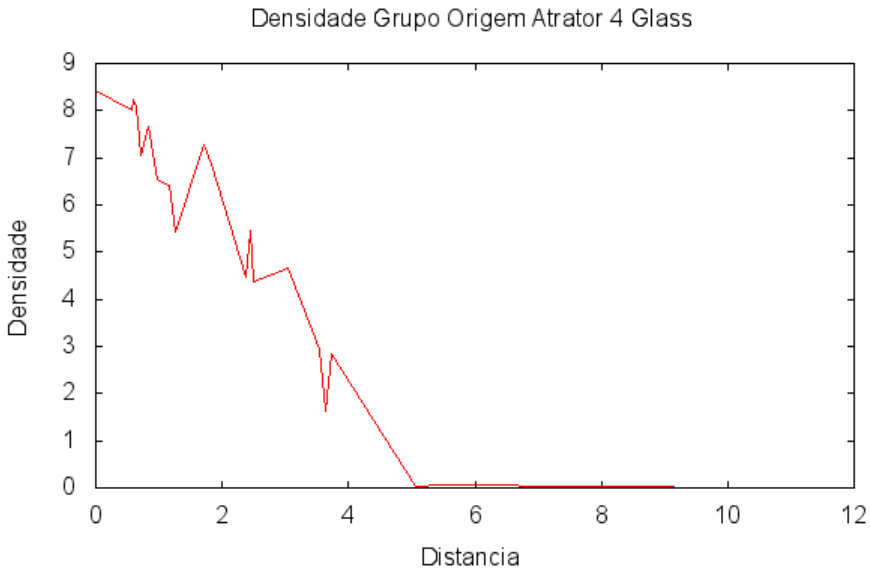

Figura 7.40. Gráfico comparativo da densidade de influência do agrupamento C4 em relação a densidade de influência do atrator A4 encontrado

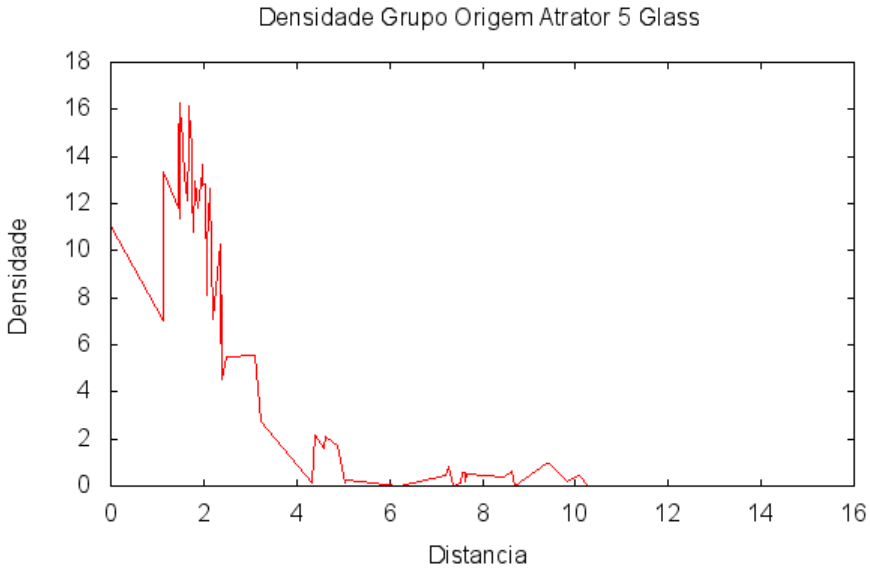

Figura 7.41. Gráfico comparativo da densidade de influência do agrupamento C5 em relação a densidade de influência do atrator A5 encontrado

Olhando novamente o gráfico 7.41 o atrator A5 relembra que sua posição, no gráfico 7.29, é distante da região de alta densidade. Este é o atrator de menor densidade em seu agrupamento, comparado com os outros atratores, o que demonstra que caso este estivesse em uma região mais próxima, possivelmente seu valor de densidade de influência se tornaria melhor. Quanto aos atratores A0 e A2 apesar de seus resultados não serem o ótimo estes estão próximos, o que os torna satisfatórios para o estudo.

Como é possível notar, a densidade de alguns pontos atratores encontrados não foram as melhores dos agrupamentos utilizados. Mas mesmo que isto tenha ocorrido, é possível notar que os resultados encontrados foram próximos aos pontos de maior densidade. Estes resultados querem dizer que com uma busca mais refinada, com diferentes valores para o numero de gerações, tamanho da população, ou até mesmo com valores diferentes de σ, é possível encontrar atratores com densidades maiores.

Nó tópico seguinte a influencia dos atratores encontrados será apresentadas. O cálculo foi feito utilizando a base como um todo, assim como no tópico seguinte foi utilizado somente os grupos a que pertence.

7.3.2.4 Gráficos de Influencia do Atrator em Relação a Base de Dados

Influencia do atrator sobre a base, com origem no atrator, nota-se que nem sempre os mais próximos sofrem maior influência. É importante ressaltar, que a influencia de um ponto sobre outro depende da distancia de um ponto sobre o outro. Caso um ponto esteja muito próximo do atrator dentro do valor σ informado, então a influencia sobre este ponto será grande. Valores fora desta região serão consideravelmente pequenos.

A Influencia é calculada utilizando todos os dados da base de dados.

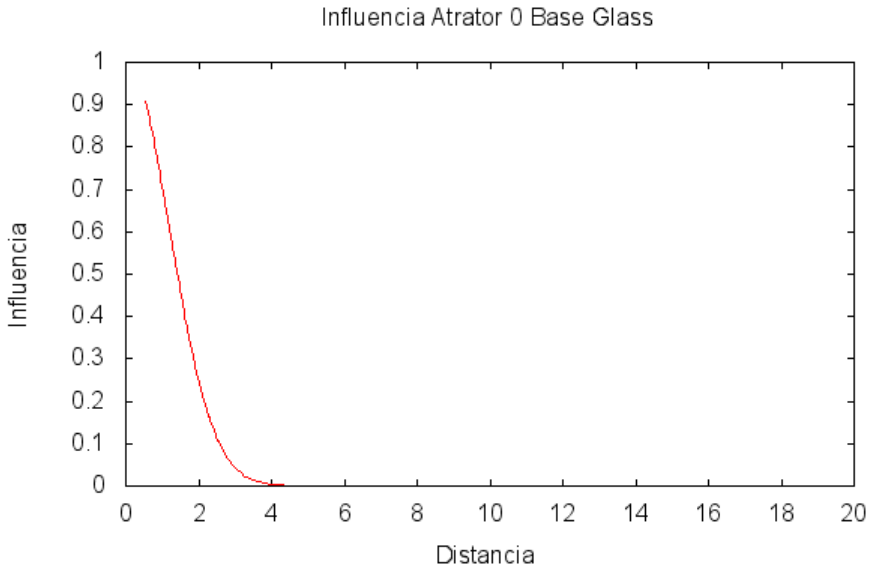

Figura 7.42. Gráfico comparativo da influencia do atrator A0 sobre a base de dados

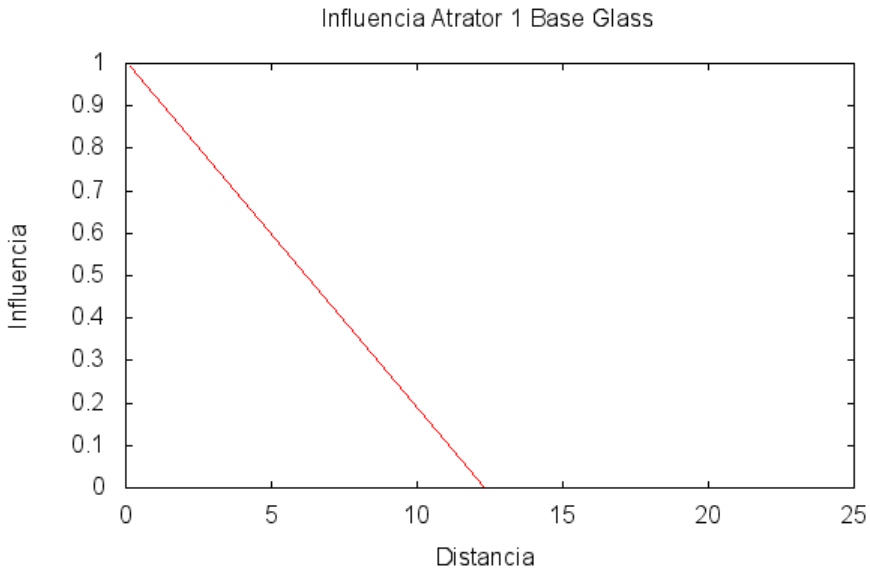

Figura 7.43. Gráfico comparativo da influencia do atrator A1 sobre a base de dados

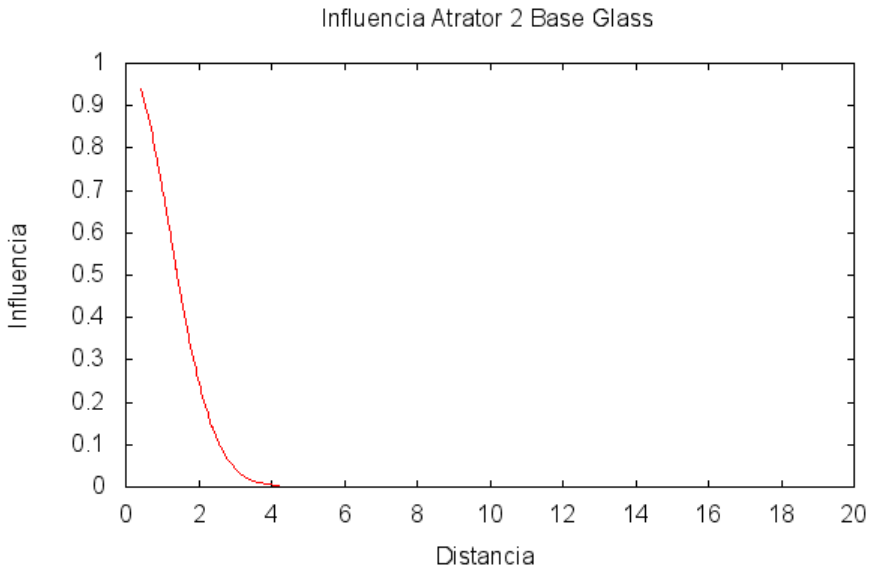

Figura 7.44. Gráfico comparativo da influencia do atrator A2 sobre a base de dados

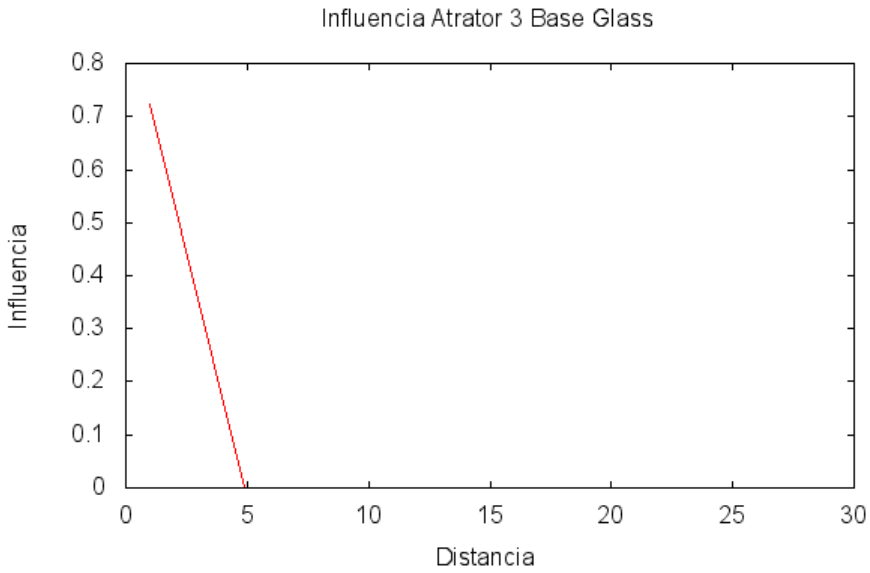

Figura 7.45. Gráfico comparativo da influencia do atrator A3 sobre a base de dados

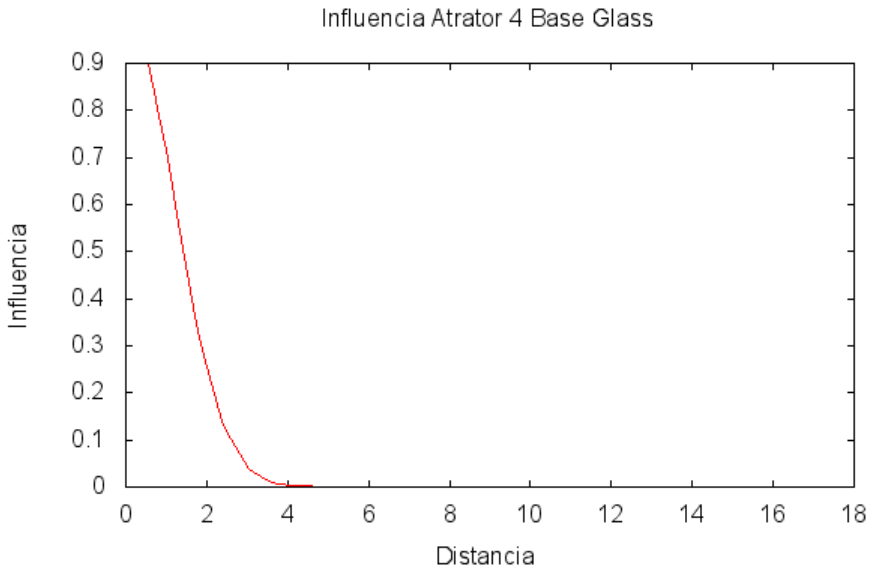

Figura 7.46. Gráfico comparativo da influencia do atrator A4 sobre a base de dados

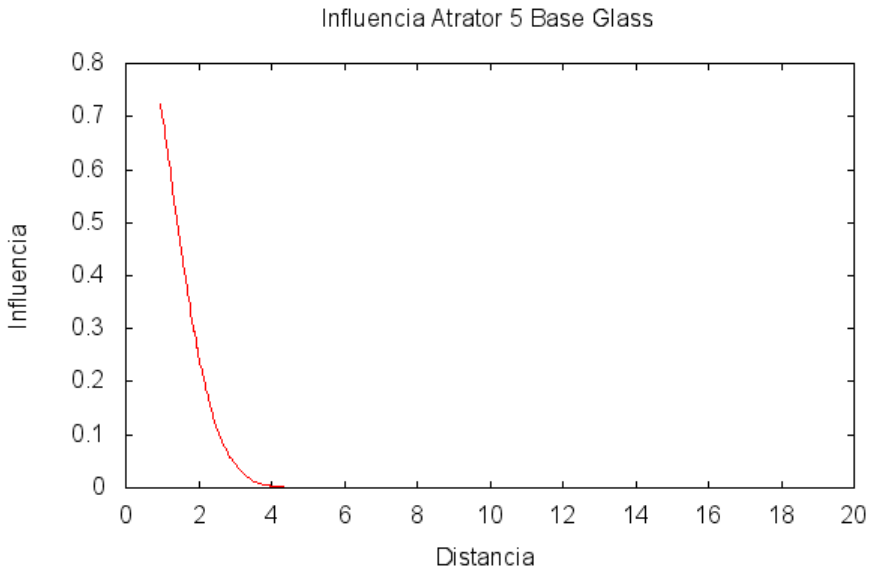

Figura 7.47. Gráfico comparativo da influencia do atrator A5 sobre a base de dados

Assim como na teoria, foi possível verificar que a influência de um atrator sobre um ponto, é maior quanto mais próximo este esteja. Os valores de influência já eram esperados, mesmo os atratores encontrados não possuindo as maiores densidades, a sua influência continua sendo constante sobre a sua região.

Á seguir os gráficos de influencia do atrator em relação a seu grupo.

7.3.2.5 Gráfico de Influencia do Atrator Sobre o Grupo a que Pertence

Os gráficos seguintes apresentam a influencia de um atrator sobre a base de dados a que pertence.

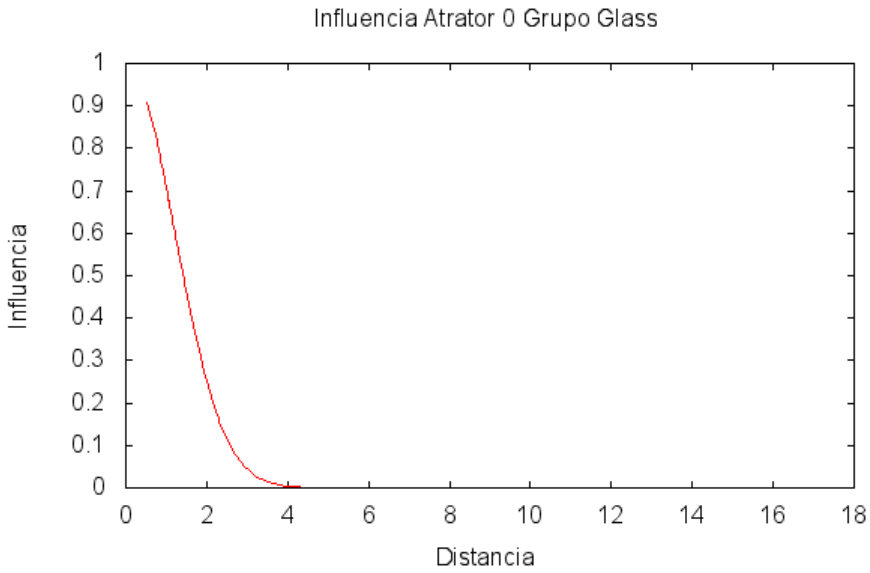

Figura 7.48. Gráfico comparativo da influencia do atrator A0 sobre o agrupamento C0 encontrado

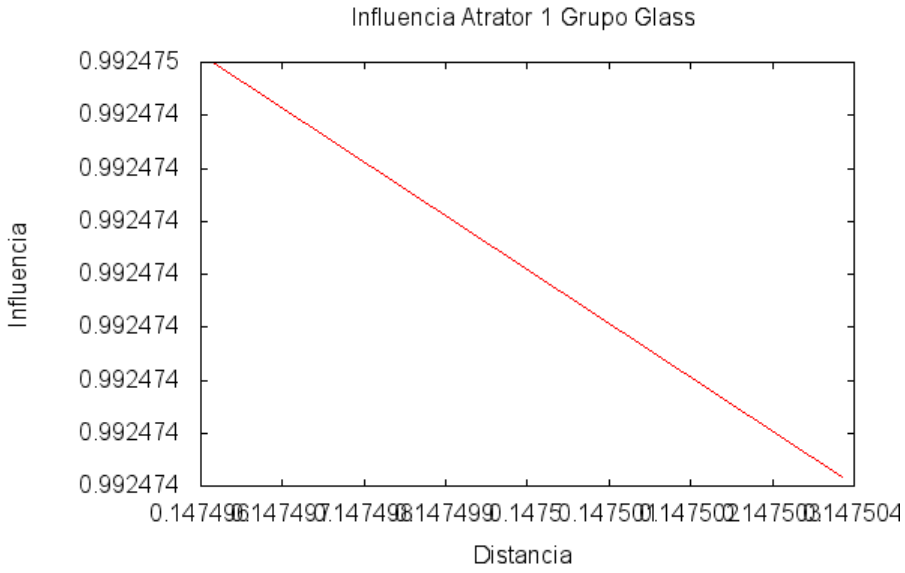

Figura 7.49. Gráfico comparativo da influencia do atrator A1 sobre o agrupamento C1 encontrado

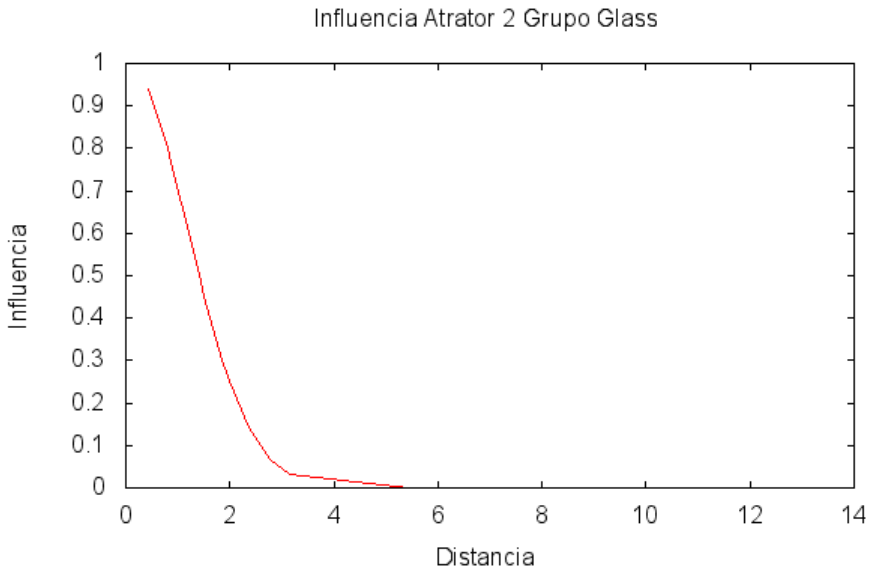

Figura 7.50. Gráfico comparativo da influencia do atrator A2 sobre o agrupamento C2 encontrado

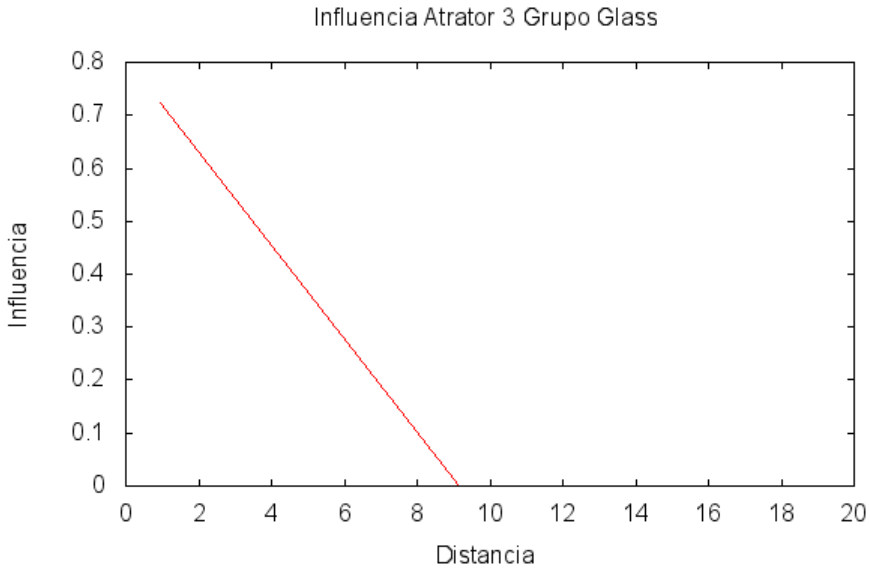

Figura 7.51. Gráfico comparativo da influencia do atrator A3 sobre o agrupamento C3 encontrado

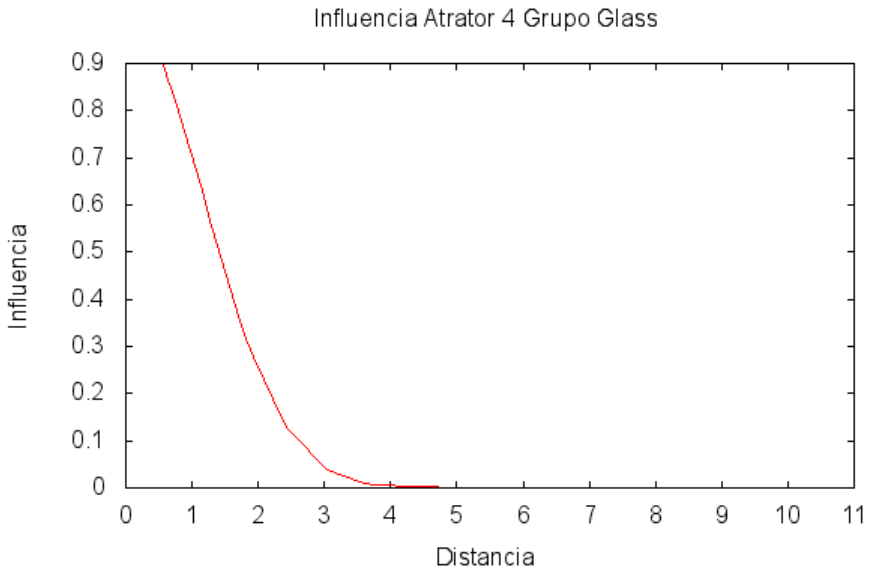

Figura 7.52. Gráfico comparativo da influencia do atrator A4 sobre o agrupamento C4 encontrado

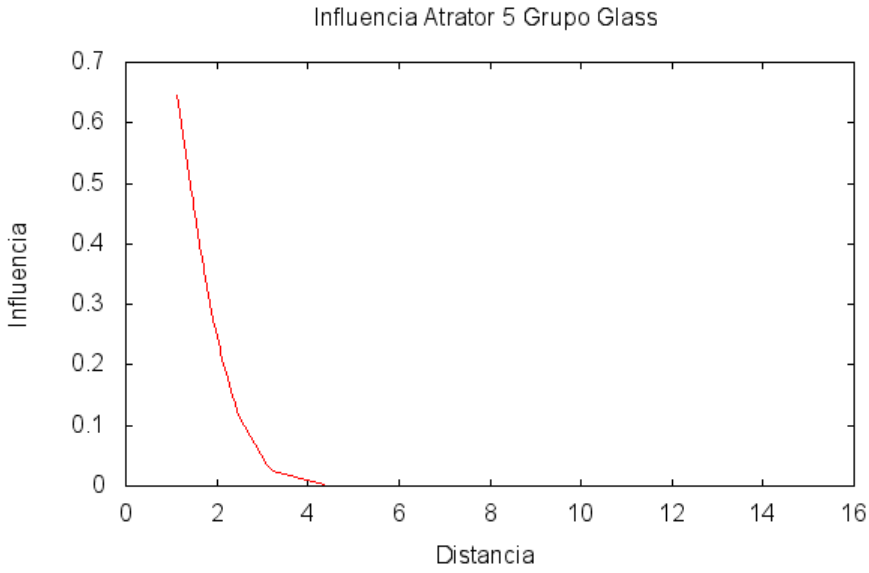

Figura 7.53. Gráfico comparativo da influencia do atrator A4 sobre o agrupamento C4 encontrado

Assim como os resultados apresentados nos gráficos comparativos da influência dos atratores sobre a base de dados, os resultados da influência dos atratores sobre a base de dados também foram os esperados.

Tabela 7.6. Acumulação de pontos na região de influência em unidades de distância.

Segmentos de Área de Influência	Número de Pontos					
	A0	A1	A2	A3	A4	A5
Até 2.00	61	2	11	1	12	23
Até 4.00	91		14	1	19	35
Até 7.00	101		16	1	21	43
Até 13.00	103		18	2	22	60
Acima de 13.00	106			4		62

Nota-se que ambos atratores, exceto o A3, possuem grande concentração de pontos em sua proximidade. Estes resultados demonstram que estes atratores possuem alto grau de atratividade em suas regiões, mesmo que suas densidades de influência não sendo as maiores encontradas nos grupos.

Esta seção apresentou alguns resultados obtidos com o algoritmo de agrupamento AGCBDG, assim como apresentou os resultados dos atratores encontrados pelo método de calculo de atratores utilizando algoritmo genético.

7.4 ANÁLISE DOS RESULTADOS OBTIDOS

Analisando os resultados apresentados neste capítulo, é possível verificar que para que os atratores consigam valores de densidade ótimos é necessário que este esteja próximo a regiões de alta densidade. Buscando esta região de alta densidade o algoritmo de agrupamento conseguiu agrupar os dados nestas regiões.

Com as regiões definidas, a busca dos pontos atratores foi realizada. Os resultados encontrados sobre a base de dados Iris são considerados ótimos, pois todos os atratores encontrados possuem a densidade de influência máxima em comparação aos pontos do agrupamento. Os resultados encontrados sobre a base de dados Glass foram considerados satisfatórios, já que para alguns atratores os resultados encontrados foram ótimos. Baseando-se nos resultados encontrados na base de dados Iris é possível analisar que podem ser encontrados melhores resultados com uma configuração melhor dos parâmetros.

Com os atratores encontrados é possível calcular a sua influência sobre a região a que pertence. Com esta influência é possível verificar o seu grau de atratividade sobre os pontos de dados. Já que o atrator age como um ponto de atração de densidade de pontos de dados.

8 CONCLUSÃO

O algoritmo de clusterização AGABDG proposto neste trabalho é utilizado como auxílio para o método de cálculo dos pontos atratores. Apesar de seu agrupamento ser de grande importância como parâmetro para os atratores, basicamente pode ser utilizado qualquer algoritmo, desde que este leve em consideração o agrupamento baseado em densidade, com o intuito de não particionar áreas de alta densidade. Pois com a divisão da região densa pode ocorrer de o atrator de densidade da região não seja calculado corretamente.

O método de cálculo de atratores de densidade utilizando algoritmos genéticos foi proposto com o intuito de estudar pontos na base de dados que podem ser considerados de grande atração para a densidade de pontos. Devido à sua localização, onde há uma maior concentração de pontos, e devido ao fato de este possuir a maior densidade de influencia sobre a sua região, o atrator pode ser considerado como um ponto de alto grau de atratividade para os outros.

Como trabalhos futuros, é possível estudar o cálculo dos atratores de densidade utilizando outros algoritmos de agrupamento, e assim analisar a densidade destes atratores sobre os grupos. Também podem ser realizados estudos utilizando outras Funções de Influencia, não somente a gaussiana como proposto neste trabalho.

Melhorias serão feitas no algoritmo de agrupamento, principalmente no que cabe à sua função de avaliação, para que este possa melhorar o seu agrupamento.

É concluído que os resultados dos atratores encontrados são bons, exceto com os atratores da base de dados Glass, mas com testes posteriores, com diferentes parâmetros, estes resultados podem melhorar.

Testes com aplicações do mundo real para os atratores podem ser feitos.

REFERENCIAS

ALPAYDM, E. ***Introduction to Machine Learning***. The MIT Press, 2004

ALVO, M. & PAN, J. **A *General Theory of Hypothesis Testing Based on Rankings***. *Journal of Statistical Planning and Inference*. Vol. 61, Issue 2, Pg. 219-248, 1997.

ANSCOMBE, F. J. ***Rejection of outliers***. *Technometrics*. Vol. 2, pg. 123–147, 1960.

BAKER, J. E. ***Adaptative Selection Methods for Genetic Algorithms***. In J. J. Grefenstette, ed., Proceedigns of the First International Conference on Genetic Algorithms and Their Applications. Erlbaum, 1985.

BEASLEY, D.; BULL, D. R.; MARTIN, R. R. ***An Overview of Genetic Algorithms: Part 1 - Fundamentals***. Inter-University Committee on Computing. University Computing, UK, 1993a.

BEASLEY, D.; BULL, D. R.; MARTIN, R. R. ***An Overview of Genetic Algorithms: Part 2 - Research Topics***. UCISA. University Computing, UK, 1993b.

BERKHIN, P. ***Survey of Clustering Data Mining Techniques***. Accrue Software, Inc., San Jose, CA, 2002.

BLASS, A. & MITAVSKIY. ***NP-Completeness of Deciding Binary Encodability***. Foundations of Genetic Algorithms 2005. Lecture Notes in Computer Science 3469. Pg. 58-74, 2005.

BLICKLE, T. & THIELE, L. ***A Comparison of Selection Schemes used in Genetic Algorithms***. Computer Engineering and Communication Networks Lab (TIK). TIK-Report No. 11, 2nd Edition, 1995.

DARWIN, C. **Origem das Espécies**. Tradução do original "On The Origin of Species: by means of Natural Selection, or the preservation of favoured races in the struggle for life", feita por Eugênio Amado. Belo Horizonte: Villa Rica, 1994.

DE JONG, K. ***Evolutionary Computation: A Unified Approach***. The MIT Press, 2006.

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

ESTER, M., KRIEGEL, H.P., SANDER, J. & XU, X. ***A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise***. In Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, 1996.

FORBES, N. ***Imitation of Life: How Biology is Inspiring Computing***. The MIT Press, 2005.

FOGEL, D. B. ***Evolutionary Computation: Toward a New Philosophy of Machine Intelligence***. Second Edition. IEEE Press, 2000

FREITAS, A. A. ***Data Mining and Knowledge Discovery with Evolutionary Algorithms***. Berlin: Springer, 264p, 2002.

GARAI, G. & CHAUDHURI, B. ***A Novel Genetic Algorithm for Automatic Clustering***, Pattern Recognition Letters, Ed. 25, pg. 173–187, 2004.

GOLDBERG, D. E. ***Genetic Algorithms in Search, Optimization and Machine Learning***. Addison Wesley Longman, Inc., 1989.

GOLDBERG, D. E. & DEB, K. ***A Comparative Analysis of Selection Schemes Used in Genetic Algorithms***. Foundations of Genetic Algorithms, 1991.

GREEN, P. E. & TULL, D. S***. Research for Marketing Decision***. Prentice-Hall, 1970.

GÜRCAN, M. N.; YARDIMCI, Y. & ÇETIN, A. E. ***Influence Function Based Gaussianity Tests for Detection of Microcalcifications in Mammogram Images***. IEEE, 1999.

HALL, L.O., OZYURT, I.B. & BEZDEK, J.C. ***Clustering With a Genetically Optimized Approach***. IEEE Trans. Evolut. Comput, 3 (2), 103–112, 1999.

HAN, J. & KAMBER, M. ***Data Mining: Concepts and Techniques*** - Second Edition. Elsevier, 2006.

HAUPT, R. L. & HAUPT, S. E. ***Pratical Genetic Algorithms***. John Wiley & Sons, Inc, 1998.

HINNEBURG, A. & KEIM, D. A. ***An Efficient Approach to Clustering in Large Mutimedia Databases with Noise***. American Association for Artificial Inteligence, 1998.

HOARE, C. A. ***Quicksort***. The Computer Journal Vol. 5 Ed. 1 Pg. 10-16, 1962.

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

HUBER, P. J. ***Robust Estimation of a Location Parameter***. Ann. Math. Statist. vol. 35, no. 1, pg. 73–101, 1964.

LORENA, L. & FURTADO, J. ***Constructive Genetic Algorithm for Clustering Problems***. Massachusetts Institute of Technology. Evolutionary Computation. v.9, n.3, 2001.

MACQUEEN, J. B. ***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, 1967.

MANLY, B. F. J. ***Multivariate Statistical Methods: A Primer. 2nd ed***., London, Chapman & Hall, 1994.

MITCHELL, M. ***An Introduction to Genetic Algorithms***. MIT Press, 1999.

MITCHELL, M. ***Genetic Algorithms: An Overview***. Adapted from An Introduction to Genetic Algorithms, Chapter 1. MIT Press. Complexity, 1 (1). pg. 31-39, 1995.

MITCHELL, T. M. ***Machine Learning***. McGraw-Hill Science/Engineering/Math, 1997.

MOTA, F. & GOMIDE, F. ***Hybrid Genetic Algorithms and Clustering***. Universidade Estadual de Campinas, Campinas – SP, Brazil, 2005.

NOVAES, U. R. **Agrupamento de Dados Através de Algorítmos Swarm**. Dissertação (Mestrado). Programa de Pós-Graduação em Engenharia. Universidade Federal do Rio de Janeiro - UFRJ., 2002.

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

PELTONEN, S. & KOUSMANEN, P. ***Output Distributional Influence Function***. IEEE Transactions on Signal Processing. Vol. 49, No. 9, pg. 1953-1960, 2001.

POLI, R. ***Tournament Selection, Iterated Coupon-Collection Problem, and Backward-Chaining Evolutionary Algorithms***. Foundations of Genetic Algorithms 2005. Lecture Notes in Computer Science 3469. Pg. 132-155, 2005.

RAKESH, A., JOHANNERS, G., DIMITRIOS, G. & PRABHAKAR, R. ***Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications***. In: Proc. of the ACM SIGMOD, p.94-105, 1999.

RECHENBERG, I. ***Cybernetic Solution Path of an Experimental Problem***. Ministry of Aviation, Royal Aircraft Establishment (U.K.), 1965.

REEVES, C. R. & ROWE, J. E. ***Genetic Algorithms - Principles and Perspectives, A Guide to Genetic Algorithm Theory***. Kluwer Academic Publishers, 2003.

RODRIGUES, M. A. P. **Problema do Caixeiro Viajante: um algoritmo para resolução de problemas de grande porte baseado em busca local dirigida**. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina – Florianópolis, 2000.

RUTKOWSKI, L. ***Computational Intelligence - Methods and Techniques***. Springer, 2008.

SARAMAGO, S. F. **Algoritmos Genéticos: A Otimização Aplicando a Teoria da Evolução**. Faculdade de Matemática. Universidade Federal de Uberlândia.

SILVA, E. B. **O Uso de Algoritmos Genéticos para Determinar Zeros de Funções Não Lineares**. Universidade Católica de Brasília. Curso de Matemática, 2006.

SILVERMAN, B. W. **Density *Estimation for Statistics and Data Analysis***. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

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

SYSWERDA, G. **Uniform C*rossover in Genetic Algorithms***. In J.D.Schaffer (Ed.) (1989) Proceedings of 3rd International Conference on Genetic Algorithms, Morgan Kaufmann, Los Altos, CA, 2-9, 1989.

TAN, P.- N., STEINBACH, M., KUMAR, V. ***Introduction do Data Mining - First Edition***. Adison-Wesley Logman Publishing co., Inc, 2005.

TEIXEIRA, O. N. **Proposta de Um Novo Algoritmo Genético Baseado na Teoria dos Jogos**. Dissertação (Mestrado em Engenharia Elétrica – Computação Aplicada) – Universidade Federal do Pará – Belém, 2005.

THEODORIDIS, S. & KOUTROUMBAS, K. ***Pattern Recognition - Second Edition***. Elsevier, Academic Press, 2003.

TUKEY, J. W. ***A Survey of Sampling from Contaminated Distributions***. Contributions to Probability and Statistics. I. Olkin. Ed. Stanford. CA: Stanford Univ. pg. 448–485, 1960.

TUKEY, J. W. ***Explaratory Data Analysis***, Addison-Wesley, 1971.

WANG W., YANG J. & MUNTZ R. ***STING: A Statistical Information Grid Approach to Spatial Data Mining***. In: Proc. of the 23rd VLDB Conf. Athens, p.186-195, 1997.

WAND, M. P. & JONES, M. C. ***Kernel Smoothing***. Chapman & Hall, 1995.

WHITLEY, D. A. ***Genetic Algorithm Tutorial***. Technical Report CS-93-103. Department of Computer Science. Colorado State University, 1993.