



**COPPE/UFRJ**

DESENVOLVIMENTO DE UMA ESTRATÉGIA DE SELEÇÃO DE  
CARACTERÍSTICAS ENCAPSULADA UTILIZANDO MODELOS DE  
APROXIMAÇÃO

Ever Pereira da Silva

Dissertação de Mestrado apresentada ao  
Programa de Pós-graduação em Engenharia  
Civil, COPPE, da Universidade Federal do Rio  
de Janeiro, como parte dos requisitos necessários  
à obtenção do título de Mestre em Engenharia  
Civil.

Orientadores: Nelson Francisco Favilla Ebecken

Carlos Cristiano Hasenclever Borges

Rio de Janeiro  
Novembro de 2009

DESENVOLVIMENTO DE UMA ESTRATÉGIA DE SELEÇÃO DE  
CARACTERÍSTICAS ENCAPSULADA UTILIZANDO MODELOS DE  
APROXIMAÇÃO

Ever Pereira da Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO  
LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA  
(COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE  
DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE  
EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

---

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

---

Prof. Carlos Cristiano Hasenclever Borges, D.Sc.

---

Prof<sup>a</sup>. Beatriz de Souza Leite Pires de Lima, D.Sc.

---

Prof. Raul Fonseca Neto, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

NOVEMBRO DE 2009

Silva, Ever Pereira

Desenvolvimento de uma Estratégia de Seleção de Características Encapsulada utilizando Modelos de Aproximação / Ever Pereira da Silva. - Rio de Janeiro: UFRJ/COPPE, 2009.

XV, 98 p.: il.; 29,7 cm.

Orientadores: Nelson Francisco Favilla Ebecken et al  
Dissertação (mestrado) - UFRJ/ COPPE/ Programa de Engenharia Civil, 2009.

Referencias Bibliográficas: p. 93-98.

1 Seleção de Características 2. Algoritmos Genéticos  
3. Classificação 4. Modelos de Aproximação. I Ebecken,  
Nelson Francisco Favilla e Borges, Carlos Cristiano  
Hasenclever II. Universidade Federal do Rio de Janeiro,  
COPPE, Programa de Engenharia Civil.  
III. Título.

Às quatro Mulheres de minha vida  
Hermínia, adorada Mãe, ausente fisicamente.  
Fátima, amada Esposa e Companheira.  
Tatiana e Ticiane, queridas Filhas

## AGRADECIMENTOS

Primeiro a Deus, por tudo e sempre.

Ao meu pai Jair, em memória, pelos ensinamentos de ética, honestidade e perseverança.

Ao Corpo Docente do PEC-COPPE, em especial ao Professor Nelson pelos conhecimentos transmitidos e paciência com minhas dúvidas durante a orientação.

Ao Cristiano, também orientador, que nos últimos meses, muito mais que orientação, prestou colaboração essencial e imensurável para a conclusão deste trabalho.

Aos Colegas do DCC/UFJF pelo apoio e incentivo de sempre, particularmente os professores Custódio, Raul, Regina, Tarcísio e Saulo que de alguma forma contribuíram durante esta caminhada.

Ao amigo Evandro, sempre prestativo, que muito ajudou nas minhas constantes desavenças com o “Word”.

Ao meu irmão e amigo Emar, minha cunhada e irmã de coração Nena, pelo incentivo e apoio incondicional de sempre.

Às minhas irmãs Lecy e Neyd pelas constantes e valiosas orações.

Embora ninguém possa voltar atrás e fazer um novo começo,  
qualquer um pode começar agora e fazer um novo fim.

(Francisco Cândido Xavier)

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

DESENVOLVIMENTO DE UMA ESTRATÉGIA DE SELEÇÃO DE  
CARACTERÍSTICAS ENCAPSULADA UTILIZANDO MODELOS DE  
APROXIMAÇÃO

Ever Pereira da Silva

Novembro/2009

Orientadores: Nelson Francisco Favilla Ebecken  
Carlos Cristiano Hasenclever Borges

Programa: Engenharia Civil

O problema de classificação trata da construção de discriminantes visando a predição de classes de uma amostragem. Na obtenção dos classificadores, todos os atributos geralmente são considerados, independentemente de sua relevância. Técnicas de seleção de características se apresentam como ferramentas que buscam aumentar a acurácia da discriminação e melhorar o entendimento dos dados avaliados. Estratégias baseadas em filtragem dos dados e, principalmente, modelos encapsulados tem apresentado resultados satisfatórios nesta tarefa. Modelos de seleção de características encapsulados buscam o subconjunto ótimo de atributos através da otimização dos resultados de predição em relação a um classificador pré-determinado, o que torna o processo computacionalmente dispendioso. Modelos de aproximação são estratégias acopladas a problemas que envolvem simulações com alto custo computacional que substituem uma parcela destas simulações por aproximações adequadas destas simulações. Gerenciado de forma eficiente, o modelo de aproximação produz resultados similares ao problema original com menor demanda computacional. Um modelo de aproximação para seleção de característica encapsulada, baseado em algoritmos evolucionistas e codificação binária é apresentado visando a redução do custo computacional, sem perda expressiva na qualidade de predição. Experimentos numéricos são realizados utilizando classificadores com propriedades diversas no encapsulamento, bem como um conjunto de bancos de dados bastante heterogêneo, para atestar a eficiência e robustez do modelo.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

DEVELOPMENT OF A WRAPPER FEATURE SELECTION STRATEGY USING APPROXIMATION MODELS.

Ever Pereira da Silva

November/2009

Advisors: Nelson Francisco Favilla Ebecken  
Carlos Cristiano Hasenclever Borges

Department: Civil Engineering

Classification problems treats with the task of construct discriminants that predict the classes of a given sampling with efficient accuracy. Usually, all attributes or features, independent of its relevance, are considered to obtain the classifier. Feature selection techniques are tools applied on databases in order to enhance the classifier accuracy and to increase the knowledge of data behavior. Filter strategies and, mainly, wrapper models have presented satisfactory results for this task. Wrapper models search for an optimum feature subset by means of an optimization prediction process in relation to a predefined classifier, usually at a high computational cost. Approximation models are strategies coupled to problems that involves high cost simulations substituting partially the simulations by its estimation by means of an adequate approximation strategy. An efficient management of the approximation produces similar results to the original problems at low computational costs. An approximation model for wrapper feature selection is developed based on evolutionary algorithms and binary encoding objecting to preserve the prediction accuracy and to reduce computational effort. Numerical experiments are performed using classifiers with different properties for the wrapper process and a heterogeneous set of data to attest the efficiency and robustness of the proposed model.

## SUMÁRIO

CAPÍTULO 1	Introdução .....	1
CAPITULO 2	Seleção de Características.....	5
2.1.	Motivação para Estudo do Problema .....	5
2.2.	Aplicação em Problemas Reais .....	6
2.3.	Definição do Problema.....	9
2.4.	Graus de Relevância.....	9
2.5.	Características Gerais dos Métodos de Seleção.....	10
2.6.	Categorização dos Métodos de Seleção de Características .....	11
2.6.1	Modelo de Seleção em Filtro .....	12
2.6.2	Modelo de Seleção em Cápsula .....	13
2.7.	Estratégia para Seleção de Características .....	16
2.7.1	Procedimentos Gerais .....	16
2.7.2	Geração do Subconjunto .....	16
2.7.3	Avaliação do Subconjunto .....	18
2.7.4	Critério de Parada .....	19
2.7.5	Validação de Resultado .....	20
CAPITULO 3	Algoritmos genéticos (AG).....	21
3.1.	Otimização.....	21
3.2.	Evolução e Otimização.....	21
3.3.	Evolução e Algoritmos Genéticos .....	23
3.4.	Algoritmos Genéticos.....	24
3.4.1	Codificação .....	25
3.4.2	Avaliação .....	26
3.4.3	Operador de Recombinação .....	27
3.4.4	Operador de Mutação .....	27
3.4.5	Critérios de Seleção .....	28
3.4.6	Roleta Simples.....	28
3.4.7	Ordenação (Ranking).....	30
3.4.8	Torneio.....	31
3.4.9	Parâmetros Genéticos .....	31
3.4.10	Esquema Geral dos Algoritmos Genéticos .....	33
CAPITULO 4	Simulação e Estratégias de Aproximação .....	34
4.1.	Modelo de Simulação.....	34
4.2.	Modelo de Aproximação ou Substituição .....	35
4.2.1	Pela Abordagem de Construção .....	35
4.2.2	Pelo Tipo de Aproximação .....	36
4.3.	Modelos de Aproximação Baseados em Similaridade.....	37
CAPÍTULO 5	Uma Estratégia de Seleção de Características Encapsulada Utilizando Algoritmo Genético com Modelos de Aproximação .....	40

5.1.	Algoritmo Genético para Seleção de Atributos .....	42
5.1.1	Função Objetivo: .....	42
5.1.2	População Inicial - Heurística: .....	43
5.2.	Classificadores .....	44
5.2.1	Máquinas de Vetores Suporte (SVM) .....	45
5.2.2	Vizinhos mais próximos (KNN) .....	46
5.2.3	Algoritmo das K-médias (K-means) .....	46
5.3.	Estrutura de Gerenciamento do Modelo de Aproximação .....	47
5.4.1	Seleção dos indivíduos a Serem Avaliados e Aproximados .....	47
5.4.2	Atualização da Base de Dados .....	48
5.4.	Especificação, Construção e Caracterização da Base de Dados (População Auxiliar).....	48
5.5.	Modelo de Aproximação .....	49
5.6.1	Aproximação pelos Vizinhos mais Próximos (VMP) .....	50
5.6.2	Aproximação por Alelo (ou Atributo) .....	50
5.6.	Algoritmo Implementado .....	54
5.7.	Aptidão por Alelo (Atributo) e Epistasia .....	55
CAPITULO 6 Experimentos Numéricos.....		60
6.1.	Recursos Computacionais .....	60
6.2.	Bases de Dados .....	61
6.3.	Experimentos .....	61
6.3.1	Bases de Dados Sonar, Breast e Ionosphere .....	61
6.3.2	Outras Bases de Dados .....	69
6.4.	Experimentos Adicionais .....	79
CAPÍTULO 7 Conclusões e Perspectivas .....		90
REFERÊNCIAS BIBLIOGRÁFICAS .....		93

## Lista de figuras

Figura 2.1 - Algoritmo do modelo em filtro	14
Figura 2.2 - Algoritmo do modelo em cápsula	15
Figura 2.3 - Algoritmos de Seleção de Características	16
Figura 3.1 - Basilosauros	22
Figura 3.2 - Nadadeira dos Tursiops	22
Figura 3.3 - Gráfico de seleção por roleta simples	29
Figura 3.4 - Gráfico de seleção por SUS	30
Figura 3.5 - Gráfico de seleção por <i>ranking</i>	31
Figura 3.6 - Fluxograma dos Algoritmos Genéticos	33
Figura 4.1 - Ilustração do procedimento de Semelhança de Aptidão	38
Figura 5.1 - Representação do cromossomo	40
Figura 5.2 - Conjunto de esquemas da função Royal Road	56
Figura 5.3 - Função MaxUns: população avaliada	57
Figura 5.4 - Função MaxUns: população aproximada	57
Figura 5.5 - Função Royal Road: população avaliada	58
Figura 5.6 - Função Royal Road: população aproximada	59
Figura 6.1 - Aptidão por geração - População Auxiliar - SONAR	64
Figura 6.2 - Percentual de características por geração - População Auxiliar - SONAR	64
Figura 6.3 - Aptidão por geração - População - SONAR	64
Figura 6.4 - Percentual de características por geração - População - SONAR	64
Figura 6.5 - Erro médio percentual por geração - SONAR	64
Figura 6.6 - Aptidão por geração - População Auxiliar - BREAST	66
Figura 6.7 - Percentual de características por geração - População Auxiliar - BREAST	66
Figura 6.8 - Aptidão por geração - População - BREAST	66
Figura 6.9 - Percentual de características por geração - População - BREAST	66
Figura 6.10 - Erro médio percentual por geração - BREAST	66
Figura 6.11 - Aptidão por geração - População Auxiliar - IONOSPHERE	68
Figura 6.12 - Percentual de características por geração - População Auxiliar - IONOSPHERE	68

Figura 6.13 - Aptidão por geração - População - IONOSPHERE	68
Figura 6.14 - Percentual de características por geração - População - IONOSPHERE	68
Figura 6.15 - Erro médio percentual por geração - IONOSPHERE	68
Figura 6.16 - Aptidão por geração - População Auxiliar - SYNTHETIC	70
Figura 6.17 - Percentual de características por geração - População Auxiliar - SYNTHETIC	70
Figura 6.18 - Aptidão por geração - População - SYNTHETIC	70
Figura 6.19 - Percentual de características por geração - População - SYNTHETIC	70
Figura 6.20 - Erro médio percentual por geração - SYNTHETIC	70
Figura 6.21 - Aptidão por geração - População Auxiliar - LEUKEMIA	72
Figura 6.22 - Percentual de características por geração - População Auxiliar - LEUKEMIA	72
Figura 6.23 - Aptidão por geração - População - LEUKEMIA	72
Figura 6.24 - Percentual de características por geração - População - LEUKEMIA	72
Figura 6.25 - Erro médio percentual por geração - LEUKEMIA	72
Figura 6.26 - Aptidão por geração - População Auxiliar - PROSTATE	74
Figura 6.27 - Percentual de características por geração - População Auxiliar - PROSTATE	74
Figura 6.28 - Percentual de aptidão por geração - População - PROSTATE	74
Figura 6.29 - Percentual de características por geração - População - PROSTATE	74
Figura 6.30 - Erro médio percentual por geração - PROSTATE	74
Figura 6.31 - Aptidão por geração - População Auxiliar - COLON	76
Figura 6.32 - Percentual de características por geração - População Auxiliar - COLON	76
Figura 6.33 - Aptidão por geração - População - COLON	76
Figura 6.34 - Percentual de características por geração - População - COLON	76
Figura 6.35 - Erro médio percentual por geração - COLON	76
Figura 6.36 - Aptidão por geração - População Auxiliar - MUSHROOM	78
Figura 6.37 - Percentual de características por geração - População Auxiliar - MUSHROOM	78
Figura 6.38 - Aptidão por geração - População - MUSHROOM	78
Figura 6.39 - Percentual de características por geração - População - MUSHROOM	78
Figura 6.40 - Erro médio percentual por geração - MUSHROOM	78

Figura 6.41 - Características selecionadas na População Auxiliar. Gerenciamento Determinístico - SONAR	80
Figura 6.42 - Características selecionadas na População Auxiliar. Gerenciamento Aleatório - SONAR	80
Figura 6.43 - Aptidão na População Auxiliar. Gerenciamento Determinístico - SONAR	80
Figura 6.44 - Aptidão na População Auxiliar. Gerenciamento Aleatório - SONAR	80
Figura 6.45 - Características selecionadas na População Auxiliar. Gerenciamento Determinístico - IONOSPHERE	81
Figura 6.46 - Características selecionadas na População Auxiliar. Gerenciamento aleatório - IONOSPHERE	81
Figura 6.47 - Aptidão na População Auxiliar. Gerenciamento Determinístico - IONOSPHERE	81
Figura 6.48 - Aptidão na População Auxiliar Gerenciamento Aleatório - IONOSPHERE	81
Figura 6.49 - Características selecionadas na População Auxiliar - Com penalidade - SONAR	84
Figura 6.50 - Características selecionadas na População - Com penalidade - SONAR	84
Figura 6.51 - Percentual de acertos na População Auxiliar - Com penalidade - SONAR	84
Figura 6.52 - Percentual de acertos na População - Com penalidade - SONAR	84
Figura 6.53 - Erro médio percentual com penalidade - SONAR	84
Figura 6.54 - Aptidão na População Auxiliar - Com penalidade - SONAR	84
Figura 6.55 - Características selecionadas na População Auxiliar - Com penalidade - BREAST	87
Figura 6.56 - Características selecionadas na População - Com penalidade - BREAST	87
Figura 6.57 - Percentual de acertos na População Auxiliar - Com penalidade - BREAST	87
Figura 6.58 - Percentual de acertos na População - Com penalidade - BREAST	87
Figura 6.59 - Erro médio percentual com penalidade - BREAST	87
Figura 6.60 - Aptidão na população Auxiliar - Com penalidade - BREAST	87
Figura 6.61 - Características selecionadas na População Auxiliar - Com penalidade - IONOSPHERE	89

Figura 6.62 - Características selecionadas na População - Com penalidade - IONOSPHERE	89
Figura 6.63 - Percentual de acerto na População Auxiliar - Com penalidade - IONOSPHERE	89
Figura 6.64 - Percentual de acerto na População - Com penalidade - IONOSPHERE	89
Figura 6.65 - Erro médio percentual com penalidade - IONOSPHERE	89
Figura 6.66 - Aptidão na População Auxiliar - Com penalidade - IONOSPHERE	89

## Lista de Tabelas

Tabela 3.1 - Seleção por roleta simples	29
Tabela 3.2 - Seleção por ranking	30
Tabela 6.1 - Bases para testes	61
Tabela 6.2 - Base de Dados Sonar 600 instâncias e 60 atributos	62
Tabela 6.3 - Base de Dados Breast 24 instâncias e 12625 atributos	65
Tabela 6.4 - Base de Dados Ionosphere 351 instâncias e 35 atributos	67
Tabela 6.5 - Base de Dados Synthetic 600 instâncias e 60 atributos	69
Tabela 6.6 - Base de Dados Leukemia 72 instâncias e 7130 atributos	71
Tabela 6.7 - Base de Dados Prostate 102 instâncias e 12600 atributos	73
Tabela 6.8 - Base de Dados Colon 62 instâncias e 2000 atributos	75
Tabela 6.9 - Base de Dados Mushroom 5644 instâncias e 98 atributos	77
Tabela 6.10 - Base Sonar. Comparação entre gerenciamentos Aleatório e Determinístico	80
Tabela 6.11 - Base Ionosphere. Comparação entre gerenciamentos Aleatório e Determinístico	81
Tabela 6.12 - Base Sonar, resultados com penalização	82
Tabela 6.13 - Base Breast, resultados com penalização	85
Tabela 6.14 - Base Ionosphere, resultados com penalização	88

# CAPÍTULO 1 Introdução

As pessoas encontram-se em um mundo discriminativo. Aonde quer que estejam, estão sempre reparando as características ou atributos do que pode ser visto e tocado. Tais observações ocorrem com um único propósito, discriminar.

Para se qualificar, dentre outras coisas, é necessário uma observação atenta do objeto a ser rotulado. Após a percepção das características expostas no objeto, o mesmo é, baseando-se no conhecimento de classificações anteriores, discriminado.

O conhecimento é construído com base em informações já obtidas e assimiladas como verdade. Considere, por exemplo, o fato de se visualizar uma pessoa que não é conhecida. O cérebro, rapidamente, baseado nas pessoas que já são de seu conhecimento, é capaz de julgar se o novo exemplar é bonito ou feio, gordo ou magro, alto ou baixo e assim por diante.

O processo de classificação é então desencadeado pela visualização do objeto a ser classificado, seguido de uma rápida avaliação de suas características ou atributos e então, discriminado entre as classes ou rótulos que se acredita verdadeiro. Uma pessoa pode ser considerada bonita por ter olhos verdes e cabelos negros, outra pode ser rotulada como magra, se possuir o peso de 75 kg e uma altura de 1.80 m.

Fica claro que as características desempenham o papel fundamental no processo de julgamento, já que são elas que verdadeiramente determinam a que classe o exemplar pertence. Porém, nem toda classificação é realizada em relação a critérios subjetivos como feio ou bonito, gordo ou magro, alto ou baixo, etc. Cientistas estão interessados em discriminar os padrões entre classes não subjetivas, como venenoso ou comestível, cancerígeno ou normal, verdadeiro ou falso e assim por diante. Desta forma, o controle de erro de classificação deve ser levado em consideração como parâmetro indicativo da qualidade do processo empregado em questão.

Erros de classificação ocorrem quando equivocadamente atribui-se um exemplar pertencente a uma determinada classe à outra. Considere duas classes C1 e C2. Se um dado exemplar pertencer à classe C1 e, baseado na análise de suas características, o atribuirmos à classe C2 diz-se que ocorreu um erro de classificação. O erro também ocorre se o exemplar pertencer à classe C2 e equivocadamente for classificado como pertencente à classe C1.

Não existe sentido em se avaliar erros de classificação para critérios subjetivos, mas em problemas reais, como os citados acima, é extremamente importante o conhecimento sobre a quantidade de erro produzido. De maneira geral, quanto menor for o nível dos erros obtidos, melhor será o processo de classificação desenvolvido com uma maior capacidade de generalização.

Sem dúvida, o cérebro humano não é o único sistema capaz de realizar classificações. Com o uso da computação foi possível a construção de máquinas capazes de classificar os padrões a elas apresentados. Tais máquinas receberam o nome de classificadores.

Os classificadores recebem como entrada todas as características do exemplar a ser discriminado e retornam como saída, a classe a qual o objeto pertence. Para a viabilização deste processo, entretanto, todas as características bem como as classes existentes são codificadas em formato numérico, para serem, enfim, interpretadas pelo classificador.

Diversos modelos de classificadores se apresentam, diferenciando-se entre si pela estratégia de construção do discriminante, forma de utilização do conjunto determinado para o seu treinamento, estratégia para generalização entre outras.

Porém, pode-se dizer que qualquer modelo usado para a obtenção de um classificador, independente de qual seja, é sensível às características das instâncias determinadas para construção do discriminante. Nem sempre o uso de todos os atributos disponíveis produz resultados satisfatórios para esta classe de problema. O mesmo se pode dizer do processo que ocorre no cérebro humano, já que este é capaz de selecionar rapidamente certos atributos que melhor classificarão o dado visualizado.

Considere, por exemplo, a tarefa de se discriminar entre sexos uma determinada pessoa. O problema é de solução trivial até mesmo para uma criança. Isso ocorre porque ela selecionará certas características que melhor discriminam as pessoas entre as classes (no caso masculino e feminino) e ignorará por completo as outras. Agindo dessa maneira ela será capaz de classificar basicamente sem erro os exemplos a ela apresentados.

De maneira oposta, os classificadores baseiam-se em todas as características dos padrões apresentados, para só então fornecer uma resposta. Assim, o processo de construção do classificador é mais lento e o número de erros relacionado pode crescer na medida em que nem todos os atributos são relevantes para a tarefa de discriminação podendo até mesmo serem nocivos ao processo. Esse fato é de fácil compreensão

quando se imagina que um classificador para efetuar a discriminação citada, utiliza todos os atributos relativos a uma pessoa, como cor dos olhos, cor dos cabelos, cor da pele, altura, peso, etc.

O controle da quantidade de erro de classificação é, como citado, essencial em problemas reais. Portanto, uma das formas de se melhorar o desempenho de um classificador, baseia-se em construir meios de selecionar as características que mais influenciam no processo de classificação ou discriminação. Técnicas de filtragem dos atributos, baseadas em medidas intrínsecas do conjunto de dados, se apresentam como uma possível estratégia para selecionar as características que indiquem uma maior relevância.

Por outro lado, os atributos normalmente têm uma relação de dependência muito forte com o classificador a ser usado. Como exemplo, imagine uma pessoa normal observando outras e discriminando-as como masculino ou feminino. O atributo “timbre de voz” pode ser considerado de pouca relevância. Mas se o observador é cego, esta característica é muito relevante e se for surdo, irrelevante.

Desta forma, técnicas que selecionam características juntamente com o classificador, conhecidas como modelos encapsulados, tendem a apresentar resultados mais eficientes em relação aos modelos de seleção em filtro. Porém, o custo computacional é, em geral, bastante elevado, pois cada possível subconjunto de características candidatas a seleção, deve ser avaliado por meio da construção do classificador correspondente. Em alguns casos, o processo de seleção encapsulado torna-se inviável, dependendo do banco de dados em questão bem como do classificador utilizado.

Modelos de aproximação são técnicas que visam substituir simulações necessárias na resolução de um determinado problema por aproximações escolhidas e implementadas adequadamente. O principal objetivo é a redução do esforço computacional inerente ao processo de simulação. Além da redução do custo computacional, o sucesso de um modelo de aproximação é também definido pela qualidade do resultado obtido em relação ao problema original. A definição de um gerenciamento eficaz, que controle adequadamente a base de dados de referência para a aproximação, determine adequadamente o balanceamento entre qual e quantas candidatas a solução devem ser aproximadas ou submetidas à simulação bem como o método utilizado para avaliar a aproximação são fatores cruciais para um bom desempenho do modelo.

Um modelo de aproximação para seleção de características encapsulado, baseado em algoritmo evolucionista e codificação binária é apresentado visando a redução do custo computacional, por meio da substituição de simulações - construção de classificadores - por modelo de aproximações específicos para esta classe de problemas. Tais modelos devem substituir a construção dos classificadores para os subconjuntos de atributos selecionados de forma que a perda na qualidade de predição, devido a aproximação, não comprometa a busca pelo conjunto ótimo de características.

A seguir, descreve-se como se dará o desenvolvimento deste trabalho.

No capítulo 2, apresentam-se as principais técnicas utilizadas para seleção de características, tanto para os modelos em filtro quanto para os modelos encapsulados. No capítulo 3, as principais propriedades e características dos algoritmos evolucionistas, em particular do Algoritmo Genético, são descritas visando apresentar o embasamento para o entendimento do modelo de aproximação que será proposto. O capítulo 4 discorrerá sobre os processos de simulação e os principais modelos de aproximação. No capítulo 5, será apresentado o modelo de aproximação proposto, construído em conjunto com um algoritmo evolucionista, para aplicação no problema de seleção de característica encapsulada. No capítulo 6, experimentos numéricos são realizados aplicando a metodologia desenvolvida em diversas bases de dados consideradas heterogêneas, inclusive em bases de expressão genética em *microarray*. Finalmente, no capítulo 7, são apresentadas as principais conclusões obtidas e sugestões para desenvolvimentos futuros.

## **CAPITULO 2 Seleção de Características**

O processo de seleção de características conhecido também como seleção de atributos tem sido um tradicional tópico de pesquisa desde os anos 70 (MOCCIARDI, 1971). A técnica é aplicada as mais diversas áreas como, por exemplo, reconhecimento de padrão, mineração de dados, aprendizado de máquina, programação matemática, entre outras.

Entre outras vantagens em se utilizar seleção de características pode-se citar o fato de que, após a sua aplicação, a dimensionalidade do espaço representativo do problema é reduzida. Assim, o processo remove os atributos redundantes, irrelevantes ou errados. Alguns resultados imediatos de sua aplicação são:

1. Modelo de classificação resultante mais preciso;
2. Melhoramento da qualidade dos dados;
3. Extração de conhecimento mais inteligível pelos seres humanos.
4. Obtenção de uma maior velocidade na execução dos algoritmos de aprendizado (indução);
5. Maior rapidez no processo de classificação;

### **2.1.Motivação para Estudo do Problema**

Vários problemas de classificação requerem o aprendizado de uma função de classificação apropriada. A função atribui um dado padrão de entrada a uma das finitas classes do problema. A escolha das características, atributos, ou medidas usadas para representar os padrões que serão apresentados ao classificador afetam, dentre outros aspectos:

- A precisão da função de classificação que pode ser obtida usando um algoritmo indutivo de aprendizado, como redes neurais ou árvores de decisão. As características usadas para descrever os padrões implicitamente definem uma linguagem de padrão. Se a linguagem não é expressiva o suficiente, o classificador não será capaz de capturar a informação que é necessária para efetuar a classificação. Dessa maneira, independentemente do algoritmo de

aprendizado usado, a precisão da função de classificação será limitada por essa escassez de informação;

- O tempo necessário para aprender uma função de classificação suficientemente precisa. Considere uma dada representação da função de classificação. As características usadas para descrever os padrões determinam o espaço de busca que precisa ser explorado pelo algoritmo de aprendizado. Uma abundância de características irrelevantes pode desnecessariamente aumentar o tamanho do espaço de busca, fazendo com que o tempo necessário para se aprender uma função de classificação suficientemente precisa, cresça;
- O número de exemplos necessário para se aprender uma função de classificação precisa. Quanto maior é o número de características usadas para descrever os padrões em um domínio de interesse, maior é a quantidade de exemplos necessários para se aprender uma função de classificação com uma precisão desejada;
- O custo de se executar a classificação usando a função de discriminação aprendida. Em várias aplicações práticas como diagnóstico médico, por exemplo, os padrões são descritos usando os sintomas observados, bem como os resultados de testes de diagnósticos. Testes usados para diagnóstico possuem além de diferentes custos, diferentes riscos a eles associados. Uma cirurgia exploratória é mais custosa e possui muito mais riscos que um exame de sangue, por exemplo;
- A compreensibilidade do conhecimento adquirido através do treinamento. A tarefa primária para um algoritmo de aprendizado indutivo é extrair conhecimento do conjunto de treinamento. A presença de um grande número de características, especialmente se elas são irrelevantes ou enganosas, podem fazer com que o conhecimento extraído seja de difícil compreensão para os seres humanos. Contrariamente, se as regras de classificação aprendidas são baseadas em um pequeno número de características relevantes, tais regras serão mais concisas e resumidas. Assim, elas se tornam mais fáceis de serem entendidas e aplicadas pelos seres humanos em outras situações.

## **2.2. Aplicação em Problemas Reais**

Uma das aplicações mais importantes do uso de seleção de características é a utilização da técnica com objetivo de se pré-processar os dados que serão usados em um

processo de mineração de dados - *data mining*. Existem, entretanto, muitas outras aplicações.

Em problemas reais geralmente encontram-se dificuldades relacionadas aos dados coletados, e entre elas podemos citar: elevado número de características, os atributos isolados não descrevem informações suficientes do padrão apresentado e alta dependência entre as características individuais.

Seres humanos são ineficientes em formular e entender hipóteses quando os dados coletados possuem um grande número de variáveis. Problemas demográficos, análise de dados biológicos ou classificação e categorização de textos são alguns exemplos que apresentam grandes dificuldades de resolução devido ao grande número de variáveis neles presentes. Esses mesmos problemas se tornam facilmente analisados se o espaço de representação for reduzido. A seleção de características procura reduzir a dimensionalidade priorizando as informações mais relevantes, dessa forma permitindo que os algoritmos de mineração de dados trabalhem de forma eficiente. Algumas aplicações ilustrativas do uso do processo de seleção de características são apresentadas abaixo.

### **Categorização de Texto**

Categorização de textos (LEOPOLD e KINDERMAN, 2002; NIGAM et al, 2000) é o problema de atribuir categorias pré-definidas a documentos existentes. O problema é de grande importância devido à enorme quantidade de documentos disponíveis na *World Wide Web*, *e-mails* e livrarias virtuais. A maior dificuldade na categorização está relacionada à alta dimensionalidade do espaço de características. O espaço original consiste de termos isolados (palavras e/ou frases) com altos valores numéricos até para textos de médio tamanho. Isso torna o problema proibitivo para vários algoritmos de mineração. Assim, busca-se a redução do espaço, sem, no entanto, sacrificar a precisão da categorização. Em (YANG e PEDERSON, 1997), diferentes métodos são usados e comparados com o objetivo de reduzir o espaço de características em problemas de categorização. Os resultados mostram que os algoritmos são capazes de remover de 50% a 90% dos termos (características) sem a redução da precisão de classificação.

## **Recuperação de Imagem**

Seleção de características é aplicada à recuperação de imagens (SWETS e WENG, 1995) com base no conteúdo. Nos últimos anos houve um grande aumento da quantidade de imagens disponíveis. No entanto, os acessos não podem ser feitos e informações não podem ser retiradas caso não exista uma organização que permita uma rápida navegação, busca e restauração. Recuperação com base no conteúdo (RUI, HUANG et al,1999) é uma técnica proposta para manipular o grande número de imagens hoje existentes. Ao invés das imagens serem classificadas por nomes, como acontece na categorização de textos, elas são indexadas pelo seu próprio conteúdo visual (características), como cor, forma, textura, etc. O grande problema em se fazer recuperação de imagens com base no conteúdo para grandes bases de dados continua sendo a ‘maldição da dimensionalidade’ (HASTIE, TBSSHIRANI et al, 2001). Como citado em (RUI, HUANG et al, 1999) a dimensão do espaço de características para o problema é, em geral, de ordem  $10^3$ . Redução de dimensionalidade é uma alternativa para a solução desse problema. As imagens, após o processo, são indexadas pelas principais características que as constituem.

## **Deteção de Intrusos**

Redes de computadores desempenham um papel fundamental na sociedade moderna e por isso se tornaram alvos de invasores, intrusos e criminosos. A segurança de um computador é comprometida quando um intruso o invade. Deteção de intrusos é uma maneira de proteger computadores contra invasões. Lee, Stolfo e Mok (2000) sugeriram um *framework* de mineração de dados para análise e construção de um modelo de deteção de intrusos. Os modelos analisam, primeiramente, a movimentação de informações em uma base de dados específica para o problema trabalhado. Dessa análise surgem certos padrões de comportamento dos usuários considerados normais. A seleção de características atua nessa etapa já que alguns padrões são descritos por várias características. O processo seleciona as características mais marcantes (relevantes) que determinam, sem perda de precisão, o padrão do indivíduo. O padrão de um novo usuário ou de um visitante não cadastrado é então comparado com o dos usuários normais, e de maneira geral, se houver muita diferença entre ambos os padrões de

comportamento comparados, o usuário é identificado como intruso e atitudes apropriadas são então efetuadas.

## **Análise Genômica**

Dados funcionais e estruturais provenientes da análise do genoma humano cresceram exponencialmente na última década, e com isso trouxeram grandes oportunidades e desafios para a mineração de dados. Em particular, a expressão genética em *microarray* (QUACKENBUSH, 2000) é uma tecnologia crescente que permite a visualização da expressão de centenas ou milhares de genes em um único experimento. No entanto, o número de padrões ou exemplos desses experimentos é extremamente reduzido. Em (DOAK, 1992), por exemplo, é apresentado um caso com 38 exemplos de treinamento mas com um espaço de característica de 7130. Casos como esses são comuns em biologia molecular. Experimentos têm mostrado que o processo de seleção de características gera resultados com precisão superior aos resultados obtidos com o uso total do espaço de representação. Alguns testes realizados com o algoritmo proposto e desenvolvido neste trabalho foram realizados sobre dados provenientes da biologia molecular.

### **2.3. Definição do Problema**

Existem três tipos de estratégias para a abordagem do problema de seleção de características:

O número de características  $q^*$  é dado. O objetivo do algoritmo de busca é decidir quais são as  $q^*$  características que melhor classificam o problema.

Seleção do subconjunto de características que melhor classificam o problema.

Seleção de subconjuntos de características em que existe a preocupação com a relação entre o tamanho do subconjunto e a precisão do classificador.

A implementação desenvolvida no trabalho leva em consideração a segunda abordagem acima citada.

### **2.4. Graus de Relevância**

Determinar quais características são relevantes à tarefa de aprendizado é a preocupação central dos que lidam com aprendizado de máquina. Isso se deve ao fato de

que a simples inclusão de uma característica irrelevante ou redundante pode reduzir dramaticamente o desempenho dos algoritmos.

Para determinar se uma característica é relevante ou não ao processo de aprendizado e classificação, é preciso compreender os conceitos de relevância. Existem diversas definições na literatura de aprendizado de máquina para o significado de uma característica ser “relevante”. Em (JOHN, KOHAVI et al, 1994; JOHN, 1997) são definidas duas notações para relevância (GUYON, 2001):

- Relevância Forte: Um atributo  $x_i$  é fortemente relevante se a sua remoção gera uma degradação no desempenho do Classificador.
- Relevância Fraca: Um atributo  $x_i$  é de fraca relevância se ele não for fortemente relevante e existir um subconjunto de variáveis  $V$  em que o desempenho do classificador usando  $V \cup \{x_i\}$  é superior ao desempenho do mesmo classificador utilizado somente sobre subconjunto  $V$ .

Existem ainda, características que não possuem relevância fraca e nem forte, por isso denominam-se irrelevantes. Estas não devem ser selecionadas.

## 2.5. Características Gerais dos Métodos de Seleção

A seleção de características tem como objetivo buscar pelas mais relevantes no espaço representativo do problema. Do ponto de vista de busca heurística, Blum e Langley (1997) argumentam que existem quatro questões que além de afetar a natureza da busca, são capazes de caracterizar qualquer método de seleção. São elas:

1. O ponto de partida no espaço de busca: Dependendo do ponto de partida escolhido, a direção de busca irá variar. A busca que começa com nenhuma característica selecionada e sucessivamente acrescenta outras, é chamada de “busca para frente” - *forward selection*. De outra forma, o método que começa com todas as características selecionadas e sucessivamente retira outras, é chamado de “busca para trás” - *backward selection*. Existem outros métodos que não seguem necessariamente as buscas acima citadas. Estes métodos se apresentam como uma mistura de ambas as técnicas.
2. Organização do procedimento de busca: claramente, se o número de características é muito grande, a busca exaustiva de todos os subespaços é proibitiva já que existem  $(2^N - 1)$  subconjuntos diferentes para  $N$  características.

Assim, a busca heurística é mais realista que a busca exaustiva, porém não garante uma solução ótima.

3. O critério de avaliação: a maneira pela qual os subconjuntos de características são avaliados é um problema importante. Para classificação, o subconjunto ideal é aquele que provê a maior separação dos dados. A separação de dados é normalmente calculada por algum critério de medida da distância entre as classes existentes. No algoritmo proposto e desenvolvido neste trabalho, a precisão de classificação foi utilizada para avaliar o subconjunto selecionado. Define-se precisão de classificação como sendo a porcentagem de exemplos classificados corretamente para um determinado conjunto de teste.
4. Critério de parada de busca: Durante o processo de avaliação, deseja-se parar o algoritmo quando é observado que não existem melhorias na precisão do classificador.

## **2.6. Categorização dos Métodos de Seleção de Características**

Há muito esforço em se comparar e avaliar os diferentes métodos de seleção de características, contudo, existem poucas tentativas na literatura em categorizá-los.

Siedlecki e Sklansky (1988) discutiram os métodos de seleção e os agruparam em três categorias temporais - passado, presente e futuro. O principal foco dos autores para divisão foi o uso da técnica *branch and bound* e de suas variantes.

Dash e Liu (1997) dividiram 32 métodos de seleção conhecidos em grupos. Os mesmos foram formados considerando-se o procedimento de busca e a função de avaliação dos métodos. Os procedimentos de busca dividem-se em: busca completa, heurística ou aleatória. As funções de avaliação estão divididas em funções relacionadas à distância, consistência e taxa de erro de classificação.

Uma classificação taxonômica mais ampla foi apresentada por Jain e Zongker (1997). Os métodos foram primeiramente divididos nos que utilizam técnicas de reconhecimento de padrão estatístico (RPE) e nos que se baseiam em redes neurais. A categoria de RPE foi então dividida em dois subgrupos. Estes estão relacionados com a monotonicidade da função de avaliação, ou seja, se ela é monotônica ou não monotônica.

Outra categorização se refere à complexidade em tempo dos algoritmos de seleção.

De maneira mais geral, os métodos para seleção de características podem ser colocados em dois grandes grupos: os que têm como resultado uma busca ótima e os que retornam soluções (sub) ótimas.

Pode-se ainda, abordar o problema de seleção de características como um problema de otimização, que então, se divide em otimização discreta e contínua.

Abaixo serão citadas as categorizações mais encontradas na literatura para a abordagem do problema, embora tal divisão não seja uma unanimidade entre os pesquisadores da área.

### **2.6.1 Modelo de Seleção em Filtro**

O modelo de seleção em filtro é a mais antiga forma de se abordar o problema de seleção de características. Antes que seja utilizado um algoritmo de aprendizado, a abordagem em filtro utiliza um critério de busca independente para encontrar o subconjunto de características apropriado. O modelo foi batizado de Filtro por John, Kohavi e Pflieger (1994) pelo fato de ele “filtrar” os atributos irrelevantes através de medições feitas entre os mesmos, antes que uma indução ocorra, ou seja, as características são selecionadas antes que o algoritmo de aprendizado seja executado. A seleção de características nesse caso é independente do classificador usado. A figura 2.1 adiante, apresenta a descrição do algoritmo de seleção em filtro.

A vantagem do modelo em filtro está no fato de que o mesmo não precisa ser re-aplicado para cada execução do algoritmo de treinamento. Assim, o modelo em filtro é eficiente ao se abordar problemas que possuam um espaço de características muito elevado.

Contudo, sabe-se, que na maioria das vezes, o subconjunto ótimo de características depende do algoritmo de treinamento a ele associado. Portanto, embora a abordagem apresente-se computacionalmente eficiente, ela encontra apenas soluções sub-ótimas genéricas que independem do classificador..

Como a abordagem em filtro não leva em consideração a variação de aprendizado introduzido pelo algoritmo de indução ela não é capaz de selecionar o subconjunto mais viável de características para o algoritmo de treinamento final. Por essa razão, um modelo de seleção em cápsula foi proposto.

## 2.6.2 Modelo de Seleção em Cápsula

A estratégia do modelo em cápsula se baseia no uso do algoritmo de indução para estimar o valor do subconjunto de características selecionado durante a fase de treinamento. A precisão do classificador após a seleção do subconjunto final é usada como seu valor ou mérito.

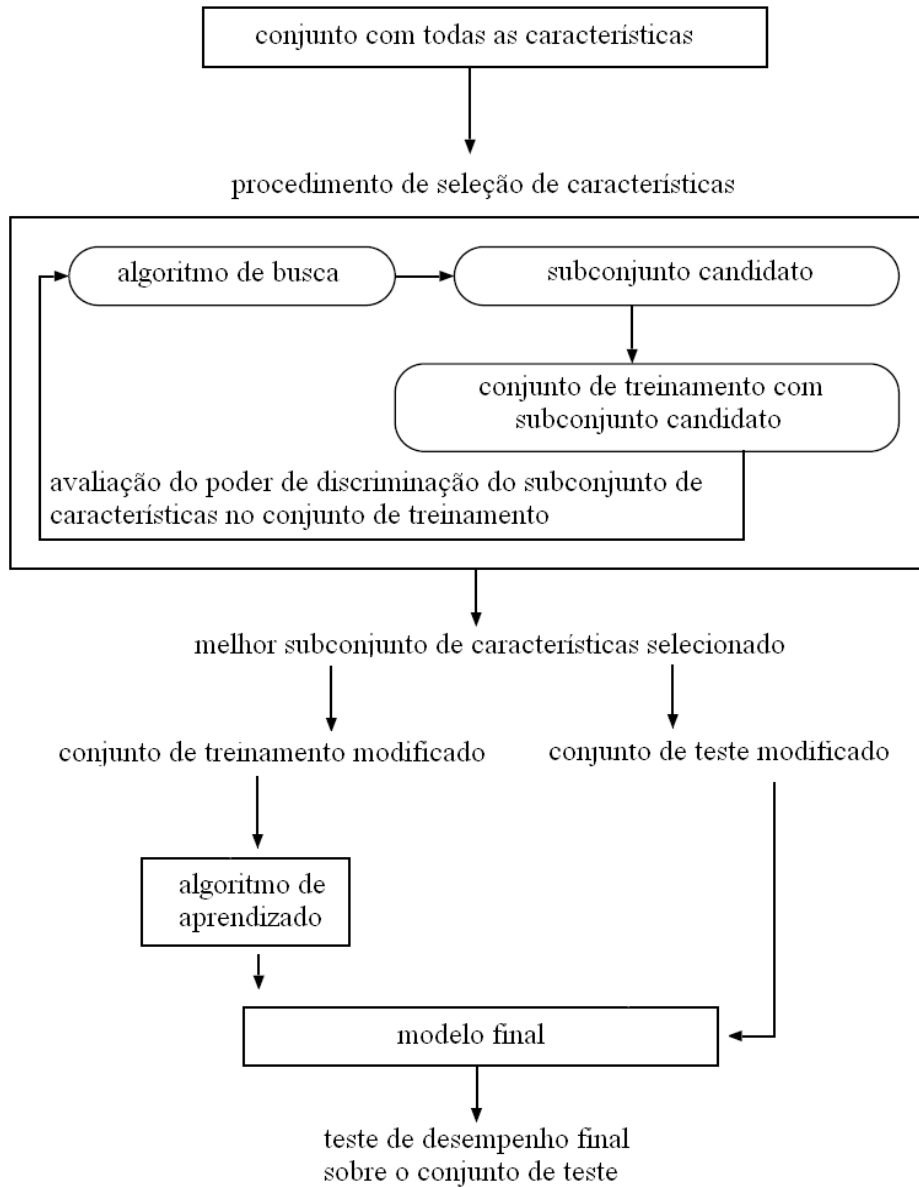
Com a abordagem em cápsula, é construído um algoritmo de treinamento específico para os dados de treinamento e a partir desse ponto, executa-se a busca pelo melhor subconjunto de características. O próprio classificador avalia o subconjunto selecionado, utilizando para isso, um conjunto de teste. Esse procedimento fornece, na maioria das vezes, melhores resultados do que quando se utiliza a abordagem em filtro. A figura 2.2 adiante, apresenta a descrição do algoritmo de seleção em cápsula.

Neste trabalho será discutido um modelo baseado em heurísticas evolucionistas, a saber, algoritmos genéticos (AG) para seleção de características. O algoritmo proposto é formulado de acordo com o modelo em cápsula. Os testes foram realizados utilizando o AG para efetuar a busca no espaço de características. Para avaliação dos subconjuntos são utilizados, separadamente, até três classificadores.

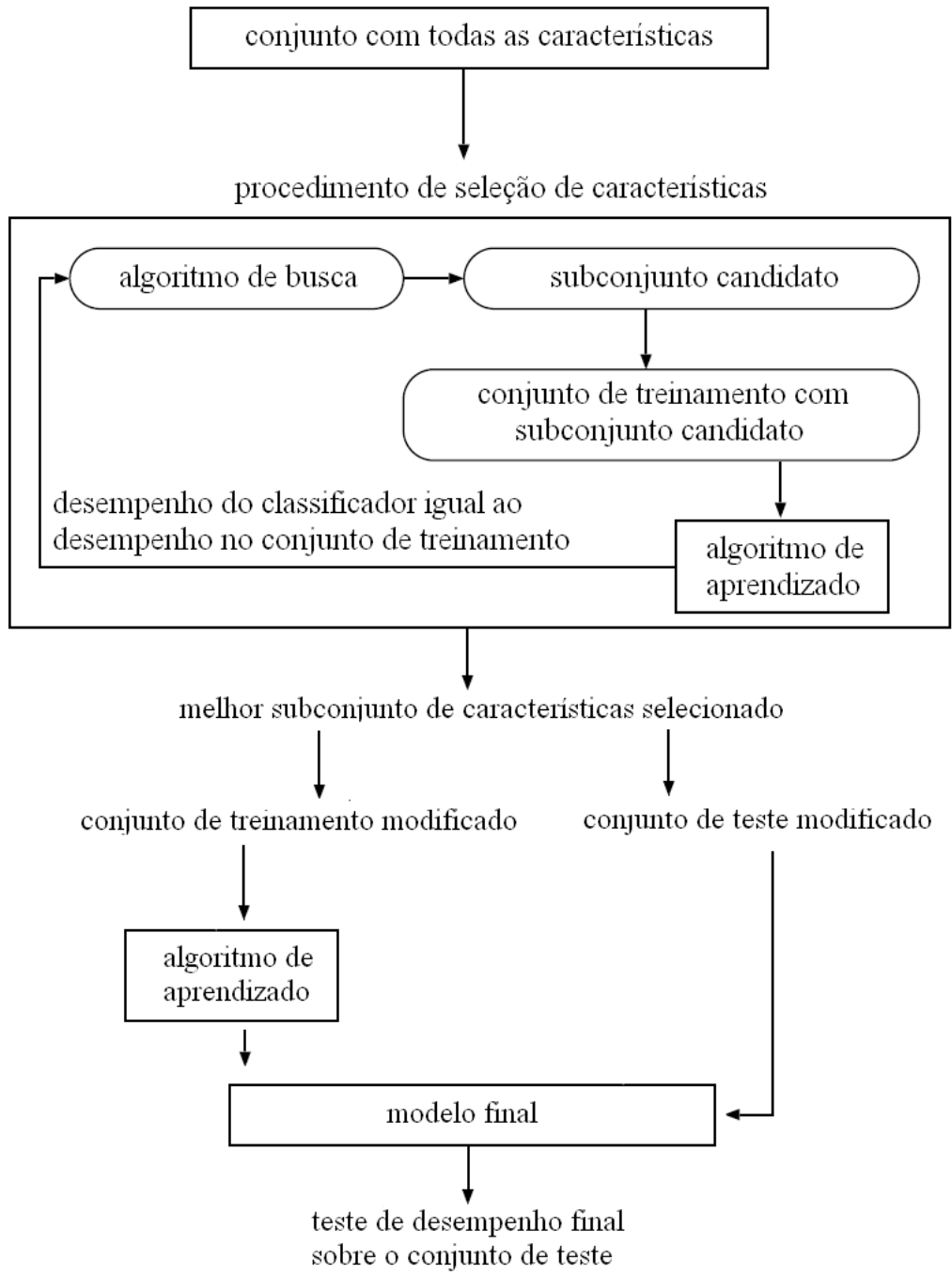
Como citado, a principal vantagem do modelo em cápsula é a dependência entre os algoritmos de seleção e de aprendizado. Essa conexão tende a obter melhores resultados do que quando a dependência é ignorada.

A principal desvantagem do método está relacionada ao tempo de execução gasto, já que este aumenta com o crescimento do conjunto de características que é, inicialmente, considerado. O tempo necessário para se executar várias instâncias do algoritmo de treinamento em um espaço que possui um número de características muito elevado pode tornar a técnica proibitiva.

A substituição de avaliações feitas pela função exata por modelos de aproximação aplicados a uma parcela de instâncias, pode melhorar sensivelmente o tempo de execução sem perda de qualidade. Esta técnica será usada neste trabalho.



**Figura 2.1** - Algoritmo do modelo em filtro



**Figura 2.2** - Algoritmo do modelo em cápsula

## 2.7. Estratégia para Seleção de Características

### 2.7.1 Procedimentos Gerais

De maneira geral, os algoritmos de seleção de características são estruturados nos seguintes passos: Geração de Subconjunto, Avaliação de Subconjunto, Critério de Parada e Validação de Resultado. A figura abaixo ilustra o processo geral dos algoritmos de seleção de características. Os passos serão detalhados em seguida.

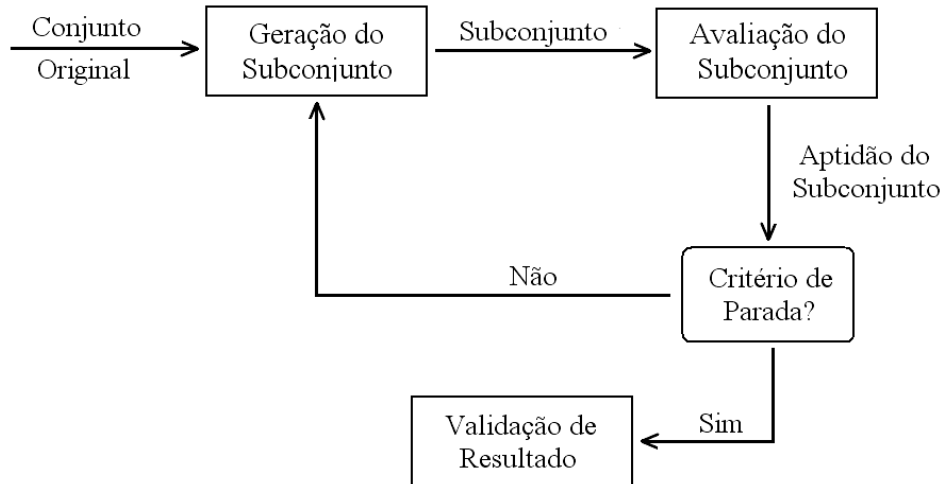


Figura 2.3 - Algoritmos de Seleção de Características

### 2.7.2 Geração do Subconjunto

A geração de subconjuntos de características é essencialmente um processo de busca heurística em que cada estado no espaço de busca corresponde a um subconjunto candidato a ser avaliado. Dois fatores principais determinam a natureza do processo.

Primeiro, é necessário determinar o(s) ponto(s) inicial(s) da busca. O algoritmo pode realizar a busca “para frente” ou “para trás” já citadas, ou realizar uma mistura de ambas as técnicas. Com objetivo de evitar que a busca fique presa em um mínimo local, a mesma pode começar em um subconjunto selecionado aleatoriamente.

Como segundo fator, tem-se a estratégia de busca. Devido ao crescimento exponencial do número de subespaços de características, diferentes abordagens para a estratégia de busca foram propostas. Elas se dividem em: busca completa, seqüencial e aleatória.

1. Busca Completa: Garante encontrar o resultado ótimo de acordo com o critério de avaliação usado. A busca exaustiva é completa, isto é, todos os subconjuntos existentes são avaliados, e com isso o resultado ótimo é encontrado. Contudo, uma busca completa não significa que ela é exaustiva. Diferentes funções heurísticas podem ser usadas para diminuir o espaço de busca sem necessariamente reduzir as chances de se encontrar a solução ótima. Assim, embora a ordem do espaço de busca ser de  $2^N$ , um número menor de subconjuntos é avaliado. Técnicas como *branch and bound* (NARENDRA e FUKUNAGA, 1977) e *bean search* (DOAK, 1992) fazem parte desse método e serão discutidas mais abaixo.
2. Busca Seqüencial: A completude do espaço de características não é explorada, portanto existe o risco de que a solução ótima não seja encontrada. Os métodos de seleção “para frente”, “para trás” ou bidirecional (LIU e MOTODA, 1998), adicionam ou removem, respectivamente, uma a uma as características do espaço. Outra alternativa é adicionar ou remover  $p$  características em uma etapa, e posteriormente, adicionar ou remover  $q$  características, sendo  $p > q$ . Algoritmos de busca seqüencial são simples de serem implementados e produzem rápidos resultados para espaços de busca com ordem de  $2^N$ .
3. Busca Aleatória: O processo começa com a escolha aleatória de um subespaço e posteriormente pode se comportar de duas maneiras distintas:
  - Seguir a busca seqüencial, dessa maneira, o método apenas inclui aleatoriedade ao método clássico de busca seqüencial. Exemplo desse método: *random start hill-climbing*.
  - Os subespaços são gerados de maneira completamente aleatória, isto é, os próximos subespaços não derivam dos anteriores mediante a uma regra fixa ou determinística, mas sim de uma forma eventual. Como exemplo pode-se citar o algoritmo Las Vegas (LIU and SETIONO, 1996).

Para todas as técnicas expostas o uso de aleatoriedade ajuda o algoritmo a escapar de mínimos locais, e a otimalidade do subconjunto selecionado depende, nesse caso, dos recursos computacionais disponíveis.

### 2.7.3 Avaliação do Subconjunto

Como mencionado anteriormente, o subconjunto selecionado precisa ser avaliado por algum critério. A aptidão de um subconjunto é sempre determinada por certo critério, ou seja, um subconjunto considerado ótimo por um critério escolhido pode não ser o ótimo de acordo com outro. De maneira geral, os critérios de avaliação são categorizados de acordo com a sua dependência em relação ao algoritmo de indução (aprendizado). Os dois grupos serão discutidos abaixo.

#### a) Critério Independente

Tipicamente, o critério de avaliação independente é usado na abordagem em filtro. O critério avalia a aptidão de uma característica (ou de um subconjunto destas) baseada nas particularidades intrínsecas do conjunto de treinamento. Alguns dos critérios independentes mais usados são: medidas de distância, medidas de dependência, medidas de informação e medidas de consistência.

**Medidas de distância** são também conhecidas como separabilidade, divergência ou medidas de discriminação. Para um problema de duas classes, uma característica  $X$  é preferível à uma característica  $Y$  se  $X$  gera um maior valor na diferença de probabilidades condicionais entre as classes abordadas. Dessa maneira, procuramos as características que melhor separam os dados entre as classes. As classes são ditas sobrepostas se o valor da diferença resultar zero.

**Medidas de informação** determinam a informação retirada de uma característica. A informação obtida usando a característica  $X$  é definida como sendo a diferença entre a incerteza a priori e a incerteza a posteriori, isto é, não selecionando e selecionando-se  $X$ . A característica  $X$  é preferível a  $Y$  se a informação obtida utilizando  $X$  é maior do que utilizando  $Y$ .

**Medidas de dependência** são conhecidas como medidas de correlação ou de similaridade. Elas medem a capacidade de se prever o valor de uma variável, utilizando os valores de outras. No processo de seleção de características para classificação, verifica-se quão uma característica está relacionada a uma determinada

classe. Uma característica X é preferível a Y se a associação entre X e uma classe C é maior que a associação entre Y e C.

**Medidas de consistência** são tipicamente diferentes das citadas acima. Essas medidas tentam encontrar um número mínimo de características que separam as classes apresentadas com a mesma uniformidade que o conjunto inteiro o faz. Uma inconsistência é definida como dois subconjuntos possuindo o mesmo número de características mas predizendo classes diferentes.

### **b) Critério Dependente**

O critério dependente é usado na abordagem em cápsula e requer um determinado algoritmo de aprendizado associado ao de busca. O algoritmo de aprendizado é aplicado ao subconjunto selecionado para avaliar a sua aptidão, e com isso, auxilia o processo de seleção apontando para uma busca mais eficiente. O processo geralmente encontra soluções melhores do que quando se utiliza o critério independente, já que a busca de características é específica para o algoritmo de aprendizado utilizado. Contudo, como citado, o algoritmo é mais custoso computacionalmente e a abordagem pode não ser viável para alguns algoritmos de aprendizado.

Por exemplo, na tarefa de classificação, a precisão de predição é normalmente usada como critério de avaliação do subconjunto selecionado. Portanto, o critério dependente é usado. Como as características selecionadas durante a fase de treinamento são as mesmas que serão utilizadas na fase de teste, e o algoritmo foi executado até se obter uma alta precisão durante a fase de treinamento, acredita-se obter bons resultados na etapa de teste. Entretanto, o esforço computacional envolvido é grande, pois o algoritmo de aprendizado é executado para todos os subconjuntos a serem avaliados.

### **2.7.4 Critério de Parada**

O critério de parada determina quando o processo de seleção de características deve ser interrompido. Alguns critérios de parada freqüentemente usados são: (a) busca completa, (b) algum dado limite (número de iterações ou uma precisão) é atingido, (c) adição ou remoção de características não produz melhores resultados, (d) um subconjunto de características suficientemente bom é selecionado segundo algum critério estabelecido.

### **2.7.5 Validação de Resultado**

Um bom critério para se validar os resultados obtidos é usar o conhecimento a priori obtido dos dados. Se as características relevantes dos dados estudados são conhecidas a priori, uma comparação direta com as características selecionadas pelo programa seria uma maneira eficaz para se validar os resultados obtidos. O conhecimento sobre características irrelevantes ou redundantes é também importante para a validação, já que se espera que estas não sejam selecionadas pelo programa.

Em problemas reais, essas informações não são conhecidas. Por isso, utiliza-se algum critério de medida indireta que relaciona o desempenho do algoritmo de aprendizado com o subconjunto de características selecionado. Por exemplo, considere que a taxa de erro do classificador foi usada como desempenho do algoritmo de aprendizado. Assim, para um dado subconjunto de características, pode-se calcular a taxa de erro e compará-la com a taxa obtida quando todo o conjunto (inicial) é utilizado. Com isso, obtém-se um excelente critério para validação e avaliação dos resultados obtidos.

## **CAPITULO 3 Algoritmos genéticos (AG)**

Baseado na teoria de Darwin (1860) tem-se uma idéia da complexidade e perfeição do processo evolutivo. Da bioquímica intrincada das células individuais à estrutura elaborada do cérebro humano, gerou-se maravilhas de inimaginável complexidade. Processos como a mutação, recombinação sexual e seleção natural proporcionam o desenvolvimento das espécies ao longo do tempo. Tentativas simplificadas de replicar a evolução natural têm sido feitas com objetivo de aplicá-las em problemas de otimização, classificação, entre outros. O conjunto de técnicas de programação evolutiva denomina-se programação genética e entre elas, se destacam os Algoritmos Genéticos.

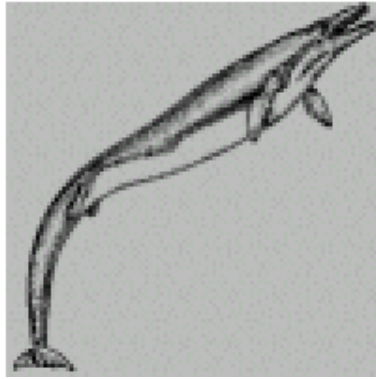
### **3.1.Otimização**

Problemas de otimização estão presentes em todas as áreas, como na economia, biologia, física e sociologia. Métodos de otimização determinísticos baseados em gradiente desempenham, sem dúvida, um grande papel na resolução destes problemas. Entretanto, alguns problemas de otimização apresentam um nível de complexidade onde tais técnicas não apresentam resultados adequados. Otimizações multimodais, com variáveis inteiras e com funções não diferenciáveis, são exemplos onde estes algoritmos são ineficientes. Outros métodos, baseados em processos probabilísticos, têm apresentado ótimos resultados nessa classe de problemas.

Estes métodos são muito eficientes, contudo, não existe a certeza de que o máximo ou mínimo global seja encontrado. A técnica somente funciona para espaços de busca reduzidos.

### **3.2.Evolução e Otimização**

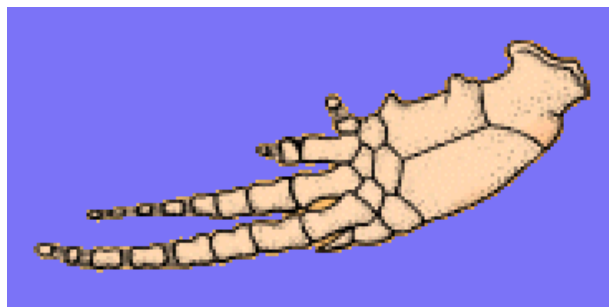
Considere um Basilosaurus, espécime existente há 45 milhões de anos, mostrado na figura 3.1:



**Figura 3.1** - Basilossauros

Os Basilossauros são considerados os protótipos das baleias. Possuíam 15 metros e cerca de cinco toneladas. Eram dotados de nadadeiras posteriores e de uma cabeça quase independente. Eles se alimentavam de pequenas presas e se moviam com movimentos ondulatórios. Seus membros inferiores eram reduzidos a pequenas nadadeiras que possuíam uma articulação no cotovelo.

O deslocamento na água é muito difícil e requer grande esforço. Assim, para se mover e manter a trajetória escolhida é necessário muita energia. Os membros anteriores dos Basilossauros não eram, portanto, realmente adaptados à natação. Para uma melhor adaptação era necessário que ocorresse um encurtamento do 'braço' bem como uma exclusão ou trava da articulação. Era indispensável, ainda, uma expansão dos dedos que iriam constituir a base da nadadeira.



**Figura 3.2** - Nadadeira dos Tursiops

A figura 3.2 mostra que os dois dedos dos atuais golfinhos (tursiops) são hipertrofiados em detrimento ao resto do membro.

Os Basilossauros eram caçadores, portanto precisavam ser rápidos e precisos. Com o passar do tempo, novas espécies apareceram com maiores dedos e menores

braços. Eles podiam se mover mais rapidamente e com maior precisão que os demais, e dessa forma, viviam por um maior período de tempo e possuíam mais descendentes.

Outras melhorias foram adquiridas considerando-se a aerodinâmica. A integração da cabeça com o corpo, mais músculos na cauda - para permitir uma maior velocidade - entre outras modificações, foram ‘conquistadas’ com o passar do tempo. Assim, produziu-se a espécie perfeitamente adaptada às restrições de um ambiente aquático.

Esse processo de adaptação, essa otimização morfológica, é tão perfeita nos dias de hoje que a similaridade entre um tubarão um golfinho e um submarino são gritantes.

Os mecanismos da evolução de Darwin estão, portanto, associados a um processo de otimização. Otimização hidrodinâmica para peixes e outros animais marinhos, aerodinâmica para pterodátiles, morcegos ou pássaros. Essa observação é a base sobre a qual se fundamentam os Algoritmos Genéticos.

### **3.3. Evolução e Algoritmos Genéticos**

No início dos anos 60, John Holland da universidade de Michigan começou seus trabalhos em algoritmos genéticos. Seu primeiro resultado foi a publicação de *Adaptation in Natural and Artificial System*. Holland (1975) possuía dois objetivos:

- (i) Melhorar o entendimento do processo de adaptação natural;
- (ii) Construir sistemas artificiais possuindo características similares aos sistemas naturais.

Sua idéia fundamental se baseava no seguinte: o reservatório genético de uma população contém, potencialmente, a solução para um dado problema de adaptação. A solução não está ‘ativa’ porque a combinação genética que a representa, está decomposta entre os vários indivíduos da população. Somente a combinação dos diferentes códigos genéticos presentes na população podem levar a solução.

Considerando o exemplo de evolução dos Basilossauros e raciocinando de maneira simplista, assume-se que o encurtamento do braço e a expansão dos dedos são controlados por apenas dois genes. Nenhum indivíduo possui esse genoma, portanto não existe ninguém completamente adaptado. Após várias gerações e recombinações genéticas direcionadas pelos processos de recombinação (*crossover*) e mutação, um indivíduo pode herdar cada gene necessário de um de seus progenitores. Dependendo dos genes recebidos, o indivíduo pode se tornar perfeitamente adaptado ao seu habitat.

Alguns métodos de otimização baseados em computação evolucionista, utilizaram a mutação como operador principal na busca do ótimo global, atingindo, desta maneira, resultados satisfatórios. No caso dos algoritmos genéticos o operador de recombinação, responsável pela troca de material genético entre os indivíduos que irão gerar novos elementos, atua como operador principal. Sua utilização induz a um processo de exploração/exploração bastante eficiente na busca de soluções ótimas ou sub-ótimas.

### **3.4.Algoritmos Genéticos**

Os algoritmos genéticos apresentam-se como uma técnica de otimização que se baseia na evolução natural. Entre suas características, o mecanismo de sobrevivência do mais apto fornece ao algoritmo a capacidade de direcionar a busca para regiões de soluções promissoras.

Desta forma, os algoritmos genéticos não necessitam realizar uma busca exaustiva para encontrar soluções viáveis (sub-ótimas ou ótimas).

Na natureza, os indivíduos mais aptos têm uma maior probabilidade de sobreviverem e reproduzirem; portanto, a próxima geração tende a ser mais apta e saudável, já que esta é descendente dos indivíduos mais adaptados da população anterior. Essa mesma idéia é aplicada nos AGs, onde parte-se de uma população inicial e seleciona-se os indivíduos mais aptos por processos probabilísticos. A prole criada a partir da combinação dos indivíduos mais aptos é então colocada em uma nova população, em sua maioria formada por indivíduos mais adequados. O processo é então repetido várias vezes para simular a idéia de evolução. O objetivo é produzir, a cada iteração, um novo conjunto de soluções melhor que o conjunto anterior. Visando proporcionar a inclusão de material genético não representado na população inicial ou perdido no processo de evolução, o operador de mutação é, geralmente, aplicado com pequena probabilidade. Os algoritmos genéticos consistem, de maneira geral, nos seguintes passos:

- Codificação;
- Avaliação;
- Seleção
- Recombinação (*crossover*)
- Mutação;
- Decodificação.

A codificação deve ser realizada de modo que cada solução seja representada de maneira única e, além disso, deve ser simples para facilitar a busca no espaço de soluções. A codificação utiliza, na maioria das vezes, o alfabeto binário, sendo o cromossomo representativo do indivíduo, uma cadeia binária.

A população inicial é gerada de forma aleatória ou mediante uma heurística pré-estabelecida. A aptidão de cada indivíduo da população é calculada; isto é, quão bem o indivíduo resolve o problema em relação aos outros. Essa aptidão é usada para calcular a probabilidade de o indivíduo ser selecionado para a próxima geração. Se o indivíduo possuir uma alta probabilidade de seleção, é mais natural a sua escolha para participar da geração da nova população. *Crossover* ou recombinação é o processo em que, geralmente, dois indivíduos são recombinaados para gerar novos indivíduos que farão parte da próxima população. O material genético que faz com que o indivíduo seja apto é repassado a sua prole através deste procedimento.

O operador de mutação seleciona aleatoriamente um ponto do cromossomo, e assim, o cromossomo ou bit selecionado é modificado.

Aplicando o processo de seleção e os operadores de recombinação e de mutação cria-se uma nova população, e a técnica é repetida até que algum critério de parada seja satisfeito. Nesse ponto, o indivíduo mais apto é decodificado no espaço.

Com objetivo de facilitar o entendimento da técnica, um exemplo simples será proposto e estudado nos sub-itens a seguir:

Considere a seguinte função:  $f = -2x^2 + 4x - 5$ , no conjunto de inteiros  $\{0,1,2,\dots,15\}$ . Busca-se maximizá-la. Pelo cálculo ou uso da força bruta, sabe-se que o valor que pertence ao intervalo e torna a função máxima é 1.

### **3.4.1 Codificação**

O processo de codificação é em geral o primeiro passo a ser dado quando são utilizados algoritmos genéticos. Quando aplicados a problemas reais, é extremamente importante encontrar uma representação para a solução que possa ser facilmente usada nos processos de evolução.

É necessária uma codificação bastante ampla, para que todas as soluções possíveis sejam representadas. A maneira mais tradicional de representação consiste na utilização de uma string binária, contendo apenas zeros e uns. Contudo, os AGs não são

restritos a esse tipo de codificação. O problema apresentado será resolvido com a codificação padrão, isto é, utilizando o alfabeto binário.

Considere o problema proposto. As soluções possíveis são apenas os números inteiros presentes no intervalo, portanto, a representação das soluções será dada pela forma binária de cada número. Por exemplo, a representação binária de 12 e 7 é 1100 e 0111 respectivamente. Note que foi adicionado o algarismo 0 na extrema direita da string que representa o número 7, e embora ele não tenha significado numérico, a alternativa foi usada para que todas as representações possuam o mesmo tamanho (quantidade de bits). Cada string, na terminologia de algoritmos genéticos, é chamada de cromossomo, e cada bit a ela pertencente denomina-se gene.

A população inicial consiste de vários cromossomos gerados aleatoriamente, isto é, o valor de cada gene em cada string assume o valor zero ou um, com distribuição de probabilidade uniforme. A quantidade de indivíduos é definida pelo tamanho da população.

### 3.4.2 Avaliação

A função de avaliação ou função objetivo é responsável pela determinação da qualidade da solução representada pelos indivíduos. Em outras palavras, a função de avaliação julga o quão ‘bom’ um indivíduo ou uma solução é para o problema tratado. Esta função está geralmente inclusa no problema.

No problema estudado, ela será simplesmente  $f = -2x^2 + 4x - 5$ , e como objetiva-se maximizá-la, soluções que tornem o valor da função maior são mais adequadas. Por exemplo, considere o valor da função de avaliação para dois números inteiros candidatos à solução, a saber, 7 e 12:

$$f(7) = -75$$

$$f(12) = -245$$

Claramente, o cromossomo 0111 que representa o indivíduo 7 é uma melhor solução que o cromossomo 1100, que representa o indivíduo 12 e, portanto, possuirá uma maior chance de reprodução. Podem-se normalizar os valores de aptidão obtidos e dessa forma, criar uma distribuição de probabilidade cumulativa. Essa probabilidade será usada no processo de seleção de cromossomos que irão gerar a próxima população.

O critério de parada é usado no processo de avaliação para decidir se a geração inteira ou se o melhor cromossomo da população atual está próximo suficiente da solução ótima. Vários critérios de parada são utilizados e podem ser usados de maneira

conjunta para que exista um maior controle da execução do algoritmo. Alguns critérios se relacionam a mínimos locais ou a solução ótima.

O critério de parada mais comumente usado se refere ao número de gerações, ou seja, após certa quantidade de iterações, o algoritmo é finalizado. Outro critério de parada está relacionado ao fato de a melhor solução não se modificar após certo número de iterações. Isso acontecerá quando o algoritmo encontrar a solução ótima, ou um ponto de ótimo local ou mesmo soluções muito próximas à ótima. Existe ainda, outro critério de parada em que o algoritmo é finalizado quando a média da aptidão da população é muito próxima à aptidão do melhor indivíduo.

### 3.4.3 Operador de Recombinação

A complexidade na aplicação da recombinação depende principalmente da forma de representação utilizada. No caso do exemplo estudado, que utiliza a codificação em cadeias binárias, a recombinação pode ser facilmente entendida e implementada.

Na chamada recombinação, escolhe-se, aleatoriamente, um ponto de corte nos cromossomos que irão gerar novos indivíduos. O material genético definido pelo ponto de corte é então trocado entre ambos. Por exemplo, usando os cromossomos

$$P_1 = 0111$$

$$P_2 = 1100$$

e supondo-se que o ponto de corte se dá após o segundo gene,

$$P_1 = 01 \mid 11$$

$$P_2 = 11 \mid 00$$

trocam-se os genes após este ponto, obtendo os novos indivíduos

$$P_1' = 01 \mid 00 = 4$$

$$P_2' = 11 \mid 11 = 15$$

Assim, de maneira geral, dois cromossomos são escolhidos aleatoriamente e então após o resultado da recombinação, suas proles são introduzidas na nova população. O processo é repetido até que a nova população esteja completa.

### 3.4.4 Operador de Mutação

A mutação é um processo completamente aleatório que encontra soluções que não poderiam ser exploradas de qualquer outra maneira. Ela é efetuada após o *crossover*

e ocorre escolhendo-se aleatoriamente um cromossomo da nova população para que este seja modificado de acordo com uma probabilidade pré-determinada.

Após a escolha do cromossomo, seleciona-se de maneira aleatória um bit do sofrerá mutação. Se o bit escolhido possuir o valor um, ele se tornará zero. Caso seja zero ele se tornará um.

Considere, por exemplo, que o cromossomo selecionado para a mutação é apresentado abaixo.

$$P_2' = 1111$$

Se o ponto escolhido para a mutação for o gene 3, após o processo o cromossomo será modificado para

$$P_2' = 1101$$

O cromossomo 1101 representa o indivíduo 13. O processo de mutação pode ter a sua complexidade aumentada dependendo da forma de representação usada.

### 3.4.5 Critérios de Seleção

O critério de seleção atua na escolha dos exemplares mais aptos para que estes, depois de passarem pelos operadores de recombinação e mutação, façam parte da nova população. Em outras palavras, a seleção age como a lei de sobrevivência dos mais aptos, atuando na escolha dos melhores indivíduos para que estes sobrevivam e reproduzam. A seguir serão apresentados os métodos de seleção mais usados.

### 3.4.6 Roleta Simples

Um método de seleção muito usado é o **Método da Roleta Simples**. Nele os indivíduos de uma geração são escolhidos para fazer parte da próxima geração através de um sorteio. Cada indivíduo tem probabilidade de ser sorteado proporcionalmente a sua adaptação,

$$P_{selex_k} = \frac{f(x_k)}{\sum_{i=1}^p f(x_i)}$$

onde  $f(x_i)$  é o valor da função de objetivo do indivíduo 'i', com  $i \in \{1, 2, 3, \dots, p\}$ , em que  $p$  é o tamanho da população.

Um exemplo simplificado é apresentado na tabela a seguir:

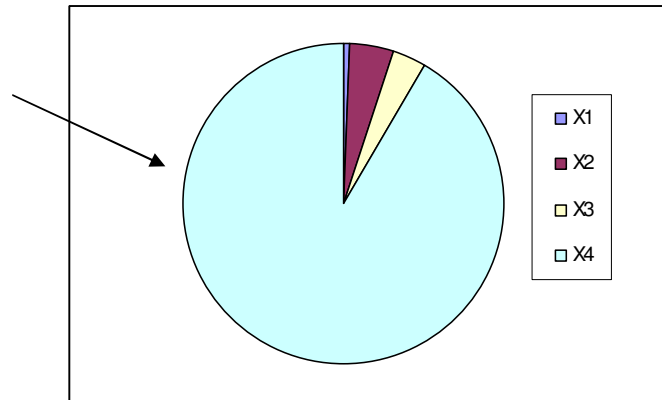
INDIVÍDUOS	APTIDÃO	PORCENTAGEM DO TOTAL
$X_4$	201	91%
$X_2$	10	5%
$X_3$	7	3%
$X_1$	2	1%
<b>Total</b>	<b>220</b>	<b>100%</b>

**Tabela 3.1** - Seleção por roleta simples

A roleta simples causa alta pressão de seleção sobre os indivíduos da população, pois o indivíduo mais adaptado, no exemplo dado, terá uma proporção de 91 para 1 de ser sorteado (selecionado para sobrevivência ou reprodução) em relação ao menos adaptado. A pressão seletiva está implicitamente relacionada com a diversidade da população.

Alta pressão seletiva tende a fazer a diversidade cair rapidamente, levando a população a convergir em poucas gerações, o que pode resultar em convergência prematura (convergência para um ponto ótimo local).

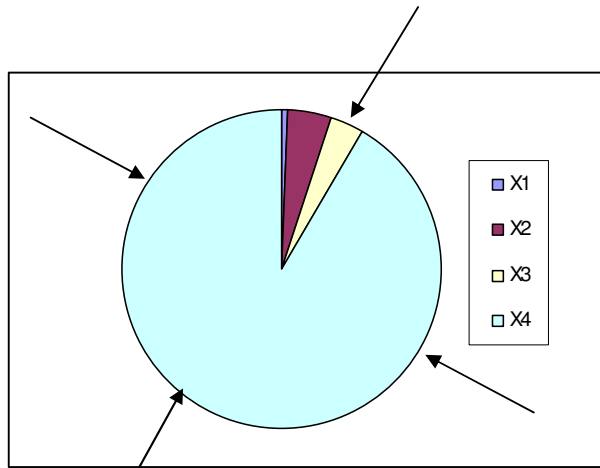
Abaixo é apresentada, em formato gráfico, as probabilidades de seleção relativas a cada indivíduo.



**Figura 3.3** - Gráfico de seleção por roleta simples

Uma variação da seleção por roleta é o método chamado Amostragem Universal Estocástica ( *Universal Stochastic Sampling* - SUS ) que será usada neste trabalho.

Semelhante à roleta, porém, para selecionar  $k$  indivíduos, utiliza  $k$  agulhas igualmente espaçadas de uma só vez, Apresenta resultados mais variantes que a roleta. A figura 3.4 abaixo ilustra este método.



**Figura 3.4** - Gráfico de seleção por SUS

O indivíduo selecionado por roleta seria ( X4), enquanto pelo critério SUS seriam selecionados (X3,X4,X4,X4)

### 3.4.7 Ordenação (Ranking)

A técnica de *ranking* (ordenação) também apresenta os indivíduos ordenados conforme suas aptidões. Contudo, no *ranking*, cada indivíduo recebe uma nota de acordo com a sua posição na ordenação. Nesse método, a probabilidade do indivíduo ser selecionado depende exclusivamente do seu ranking e não do valor de sua aptidão.

A tabela 3.2 apresenta os dados de quatro indivíduos ordenados para seleção segundo o critério de seu *ranking*:

INDIVÍDUOS	ADAPTAÇÃO	DISTÂNCIA
X <sub>4</sub>	201	40 %
X <sub>3</sub>	10	30 %
X <sub>2</sub>	7	20 %
X <sub>1</sub>	2	10 %
<b>TOTAL</b>	<b>220</b>	<b>100 %</b>

**Tabela 3.2** - Seleção por ranking

Pode-se estabelecer um comparativo entre os dois critérios citados anteriormente. O gráfico da figura 3.5 apresenta um critério de seleção por *ranking* com os mesmos dados referentes ao critério de roleta simples apresentado anteriormente. Percebe-se, nitidamente, que o critério de seleção por *ranking* não leva a uma alta pressão de seleção e, conseqüentemente, evita a uma conversão prematura.

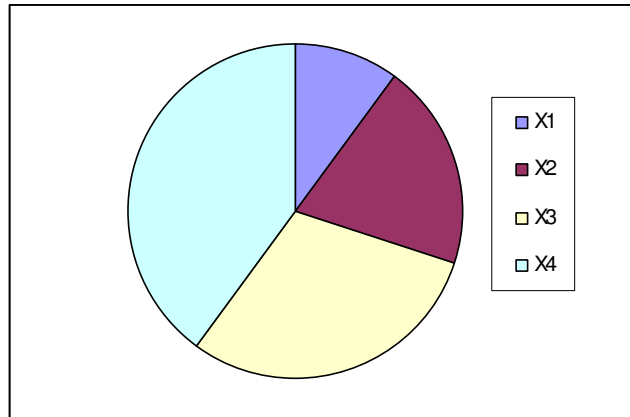


Figura 3.5 - Gráfico de seleção por *ranking*

### 3.4.8 Torneio

O critério de seleção por torneio, como o nome sugere, supõe que exista uma competição entre os indivíduos selecionados. O torneio pode ocorrer entre qualquer número de indivíduos, porém o mais comum é a competição binária. Nesse caso, são selecionados dois indivíduos de maneira aleatória e os valores de suas aptidões são comparados. O exemplar que for mais apto ganhará o torneio e será transferido para a próxima população.

O torneio permite um maior controle sobre a pressão de seleção, já que o número de indivíduos que estarão competindo pode variar durante a execução do algoritmo.

### 3.4.9 Parâmetros Genéticos

É importante analisar de que maneira alguns parâmetros influem no comportamento dos Algoritmos Genéticos, para que se possa estabelecê-los conforme as necessidades do problema e dos recursos disponíveis:

- **Tamanho da População** O tamanho da população afeta o desempenho global e a eficiência dos Algoritmos Genéticos. Com uma população pequena, a busca pode ser ineficiente, pois deste modo existe pouca diversidade genética para representar o espaço a ser explorado. Uma grande população geralmente fornece uma diversidade adequada do espaço do problema, prevenindo convergências prematuras para soluções locais ao invés de globais. No entanto, para se trabalhar com grandes populações, são necessários maiores recursos computacionais, ou que o algoritmo trabalhe por um período de tempo muito maior;

- **Taxa ou Probabilidade de Recombinação** Quanto maior for esta taxa, mais rapidamente novas estruturas serão representadas na população. Mas se esta for muito alta, estruturas com boas aptidões poderão ser retiradas mais rapidamente, assim a maior parte da população será substituída e pode ocorrer perda de estruturas de alta aptidão. Com um valor baixo, o algoritmo pode tornar-se muito lento;
- **Taxa de Mutação** Uma baixa taxa de mutação previne que uma dada posição fique estagnada em um valor, além de possibilitar que se chegue a qualquer ponto do espaço de busca. Com uma taxa muito alta a busca se torna essencialmente aleatória.
- **Taxa de Reposição** Controla a porcentagem da população que será substituída durante a próxima geração. Com um valor alto, a maior parte da população será substituída, assim pode ocorrer perda de estruturas de alta aptidão. Com um valor baixo, o algoritmo tornar-se muito lento.
- **Número Máximo de Gerações:** deve ser bem definido de forma a garantir que se tenha obtido uma busca eficiente de acordo com a diversidade genética presente na população. Geralmente é utilizado como critério de parada.

### 3.4.10 Esquema Geral dos Algoritmos Genéticos

Na figura 3.6 é apresentado um esquema que simboliza, de maneira geral, o funcionamento dos Algoritmos Genéticos.

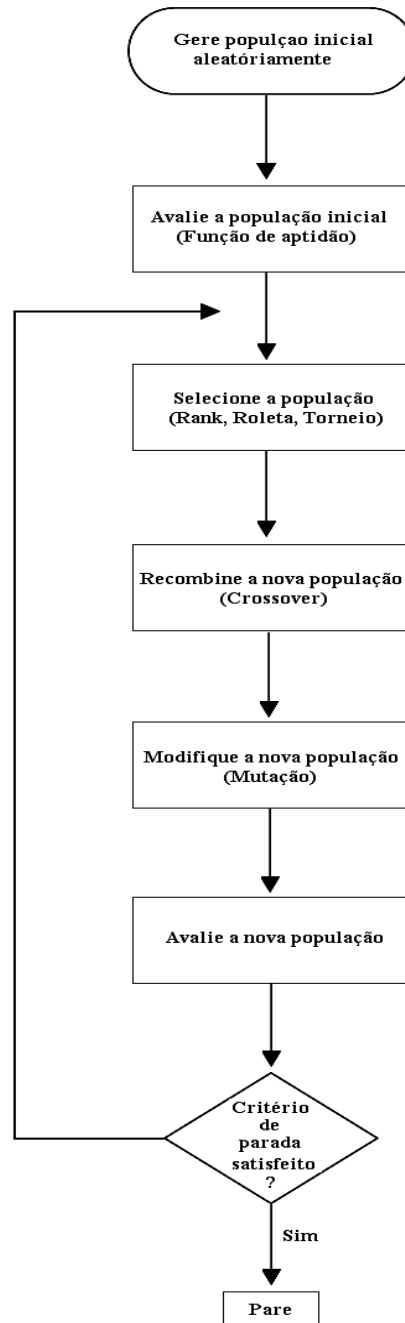


Figura 3.6 - Fluxograma dos Algoritmos Genéticos

## CAPITULO 4 Simulação e Estratégias de Aproximação

Modelos de aproximação definem estratégias que buscam obter uma solução aproximada visando um menor número de simulações, com qualidade adequada. São modelos de menor custo computacional do que os modelos de simulação. Usando-se a premissa de seleção natural, aplicada em conjunto com técnicas de computação evolucionista, pode-se encontrar melhores soluções.

Porém o uso somente destes modelos de aproximação pode apontar para resultados insatisfatórios. Desta forma, sugere-se o uso combinado dos modelos de simulação e aproximação no processo de evolução (JIN,2005). Este controle de evolução, quando usado em computação evolucionista, pode ser feito de duas maneiras distintas ou combinação entre elas:

- Controle de evolução baseado no indivíduo:

Em cada geração alguns indivíduos são avaliados com o modelo de simulação e outros com o de aproximação.

- Controle de evolução baseado nas gerações:

Algumas gerações são avaliadas com o modelo de simulação e outras com o de aproximação.

### 4.1.Modelo de Simulação

É a implementação computacional de um modelo matemático de um sistema, onde se obtém dados e medidas de desempenho através de uma simulação (CHEN et al,2006; JACOBS et al, 2004). Não são de uso geral e cada um é bastante relacionado com o problema em questão. Pode também ser visto como uma aplicação que transforma dados de entrada em saída segundo um modelo lógico ou matemático (HENDRICKX e DHAENE, 2005)

Representação de um modelo de simulação:

$\mathcal{H}$  ou  $\mathcal{H}(x)$  onde  $x \in S^n$  são variáveis de entrada,  $S^n$  é o espaço de busca e  $n$  a

dimensionalidade deste espaço.

## 4.2. Modelo de Aproximação ou Substituição

É uma implementação de menor custo computacional, aproximada do modelo de simulação. Esta aproximação pode ser uma construção baseada em uma quantidade limitada de dados ( $\mathcal{D}$ ) resultante de experimentos computacionais ou um modelo simplificado, derivado do modelo de simulação, porém com hipóteses numéricas ou físicas menos rígidas.

Representação de um modelo de aproximação

$$\mathcal{H} \approx \hat{\mathcal{H}} = \hat{\mathcal{H}}(\mathcal{D}, \alpha, \chi) \text{ onde}$$

$$\mathcal{D} = \{(\chi^i, \mathcal{H}(\chi^i)), i=1, \dots, n\}, \chi^i \in \mathcal{S}$$

$\alpha$  são os parâmetros do modelo de aproximação,  $\mathcal{S}^n$  é o espaço de variáveis de entrada,  $n$  é a dimensionalidade do espaço de variáveis de entrada,  $\chi \in \mathcal{S}^n$  são as variáveis de entrada,  $n$  é o tamanho do conjunto armazenamento  $\mathcal{D}$  e  $\hat{\mathcal{H}}(\mathcal{D}, \alpha, \chi)$  as saídas ou o resultado da aplicação do modelo de aproximação. O símbolo  $\approx$  tem a conotação de aproximadamente.

Métodos de aproximação, interpolação, construções derivadas de modelos de simulação e modelos estatísticos podem ser considerados modelos de aproximação. Dentre os diversos, podemos destacar:

- Máquinas de Vetores Suporte (KECMAN, 2001)
- Redes Neurais (KRÖSE e VAN DER SMAGT, 1993; MITCHELL, 1997; FERRARI e STENGEL, 2005)
- Funções de Base Radial (KYBIC et al, 2002; MULLUR e MESSAC, 2006; HUSSEIN et al, 2002)
- Processos Gaussianos ou Kriging (EMMERICH et al, 2006; VAN BEERS e KLEIJNEN, 2004; ULMER et al, 2003)
- Modelos Polinomiais lineares e não lineares (LESH, 1959; BLANNING, 1974; SAUNDERS et al, 1998)

Pode-se classificar os Modelos de Aproximação sob dois enfoques distintos:

### 4.2.1 Pela Abordagem de Construção

Um modelo de aproximação pode ser construído com foco no ajuste de dados ou derivado do modelo de simulação

**- Foco no ajuste de dados:**

Construído com base nos dados disponíveis de simulações anteriores, dados estes gerados especificamente para o modelo ou retirados das primeiras iterações com o método de otimização.

Esta abordagem independe do modelo de simulação, mas depende de um procedimento de ajuste de parâmetros que pode envolver a solução de um sistema linear, um processo de aprendizado ou um subprograma de otimização.

É a forma de construção mais difundida pela sua facilidade de entendimento e aplicação e será utilizada neste trabalho.

**- Foco no Modelo de Simulação:**

Construído com base no modelo original, porém utilizando hipóteses físicas menos rígidas do fenômeno que representam ou simplificação numérica do modelo de simulação.

Normalmente dependem de um especialista para validar sua construção e interpretação.

Podem ser classificados em modelos de convergência variável, resolução variável ou fidelidade física variável.

Independente do modelo da aproximação adotado, a vantagem em termos computacionais depende da relação entre o custo computacional dos modelos de simulação e substituição relacionado também com a qualidade de predição.

#### **4.2.2 Pelo Tipo de Aproximação**

São três, os tipos de aproximação estudados:

**- Aproximação do problema**

Consiste em substituir o problema original por outro semelhante, mais simples e de fácil resolução. Esta abordagem requer um conhecimento prévio do fenômeno analisado e do modelo de simulação utilizado. Do fenômeno, para avaliar quais termos podem ser retirados do modelo matemático e do modelo de simulação para modificá-lo corretamente, retirando trechos com significativo custo computacional.

**- Aproximação Funcional**

Consiste em gerar um conjunto de experimentos computacionais que possam mostrar de maneira segura, a relação entrada/saída do modelo de simulação e então

ajustar o modelo de aproximação aos dados disponíveis, usando-o em substituição ao modelo original (SARAIVA, 1997; PEREIRA, 2002 e MENDONÇA, 2004).

#### **- Aproximação Evolucionista**

Usada em Algoritmos genéticos e evolucionistas e se trata de um processo para estimar as aptidões de alguns indivíduos, baseado na similaridade deles com indivíduos avaliados anteriormente. Neste trabalho foca-se esta aproximação.

### **4.3. Modelos de Aproximação Baseados em Similaridade**

Modelos de aproximação devem ser o mais simples possível e a quantidade de parâmetros envolvidos, mínima, porém garantindo precisão satisfatória nas estimativas. Excesso de parâmetros pode tornar o modelo complexo, e o processo de ajuste aumentar o custo computacional. (BLANNING, 1975)

Modelos baseados em similaridade armazenam um conjunto de exemplos e cada vez que um novo elemento é submetido a esta classe de modelos, sua relação com os exemplos armazenados é construída com o objetivo de atribuir a este novo exemplo, um valor funcional.

Estes modelos podem ser gerados com ênfase em um dos três métodos mostrados a seguir:

#### **- Herança de Aptidão**

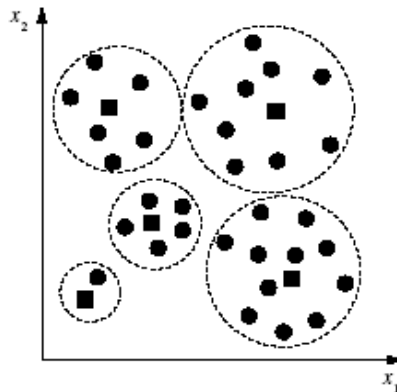
A herança de aptidão é um das técnicas de melhoramento de eficiência citadas em (SASTRY, 2002), e um tipo de aproximação evolucionista de acordo com (JIN, 2005). No caso mais simples, a aptidão de novos indivíduos é derivada da média dos valores de aptidão dos genitores. Este procedimento foi primeiramente sugerido por Smith et al (1995), que propuseram duas formas de herança: uma tomando a média e outra tomando a média ponderada da aptidão dos dois genitores.

#### **- Semelhança de Aptidão**

Na técnica de semelhança de aptidão, (JIN, 2005), os indivíduos na população são arranjados em diversos grupos. Várias técnicas de agrupamento (*clustering*) podem ser usadas para realizar esta tarefa (KIM e CHO, 2001). Então, um indivíduo é escolhido para representar cada grupo, que será avaliado pelo modelo de simulação. A escolha do indivíduo representativo pode ser feita deterministicamente ou randomicamente (MOTA e GOMIDE, 2006). Os valores da aptidão dos outros

indivíduos do mesmo grupo serão estimados de acordo com uma medida de distância em relação ao representante do grupo. Se um novo indivíduo que deve ser avaliado aproximadamente não pertence a nenhum grupo já existente, então ele é avaliado com o modelo exato.

Uma ilustração destes grupos é mostrada na Figura 4.1. Os indivíduos dentro das regiões pontilhadas pertencem a um mesmo grupo. O indivíduo mais próximo do centro do grupo, neste caso mais próximo do centro da circunferência, denotado por um quadrado, é escolhido o indivíduo representativo do grupo e será avaliado pelo modelo de simulação. Os outros indivíduos serão avaliados pelo modelo de aproximação em função da distância para o de avaliação exata. Exemplos de aplicação desta técnica podem ser encontrados em (MOTA e GOMIDE, 2006; KIM e CHO, 2001; AKBARZADEH-T et al, 2008).



**Figura 4.1:** Ilustração do procedimento de Semelhança de Aptidão. Os indivíduos dentro do círculo pontilhado pertencem a um mesmo grupo. O indivíduo representativo, denotado pelo quadrado negro, é avaliado com a função de avaliação exata (modelo de simulação). Os indivíduos restantes são avaliados pelo modelo de aproximação, e seus valores de aptidão são preditos em função da distância para o indivíduo representativo.

### **-Vizinhos mais Próximos**

Um dos modelos mais simples e utilizado para aproximação é o método do vizinho mais próximo. Considera os  $k$  indivíduos avaliados pelo modelo de simulação mais próximos ao que se busca a aproximação, por uma medida de distância previamente estabelecida. Em seguida calcula-se a média aritmética das distâncias destas  $k$  amostras. Este método é conhecido por KNN, do termo em inglês *k-Nearest Neighbor Algorithm*, onde  $k$  indica o número de vizinhos usados para a avaliação.

. Pode-se considerar, também, o algoritmo KNN utilizando a ponderação da contribuição de cada um dos  $k$  vizinhos de acordo com a distância das amostras. A idéia é associar maior contribuição às amostras mais próximas.

Outros modelos podem ser definidos para cálculo das distâncias. Por exemplo, Shepard (1968), apresentou uma abordagem usando médias ponderadas e valores funcionais, para interpolar valores em uma malha bidimensional, onde os dados estão irregularmente distribuídos.

Neste trabalho, o algoritmo KNN será utilizado em dois níveis. Como um classificador para seleção encapsulada e como técnica para o cálculo de distância no modelo de aproximação, sendo adotada a média aritmética das distâncias entre as  $k$  amostras.

Será introduzido um outro método de aproximação baseado na estrutura da população do algoritmo evolucionista, considerando a aptidão relativa dos atributos que será chamada de aproximação por atributo e detalhada no capítulo seguinte.

## CAPÍTULO 5 Uma Estratégia de Seleção de Características Encapsulada Utilizando Algoritmo Genético com Modelos de Aproximação

Neste capítulo será desenvolvido um modelo específico para aproximação em problemas de seleção de características encapsulado com algoritmos genéticos.

Os indivíduos da população de um algoritmo genético binário são codificados em cromossomos onde identificam-se os alelos que compõem os genes e os genes que irão codificar as variáveis de projeto que definem o problema a ser tratado. A figura 5.1 mostra uma representação de um cromossomo:

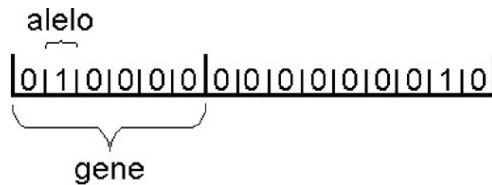


Figura 5.1 Representação do cromossomo

A codificação de um problema de seleção de características em um cromossomo é, geralmente, determinada por meio de uma cadeia binária com comprimento igual ao número de atributos existentes em cada instância. Desta forma, tem-se uma simplificação em relação ao cromossomo tradicional onde, neste caso, cada gene é formado de um simples alelo. Esta propriedade será determinante para a construção da técnica de aproximação proposta especificamente para o modelo.

A aplicação desta codificação é bastante simples. O indivíduo a ser avaliado tem suas características ativas determinadas (alelo = 1) e as demais inativas (alelo = 0), retiradas do banco de dados antes do processo de treinamento do classificador. Objetiva-se que a aplicação dos operadores genéticos no decorrer das gerações determinem o subconjunto de características que devam estar ativas e que otimizem a construção de um classificador mais eficiente, isto é, com um maior poder de generalização.

Como visto, em um método de seleção de atributos encapsulado o custo computacional é bastante alto pois para cada indivíduo da população deve-se construir um classificador específico para os atributos ativos. Exemplificando, se for utilizado

uma população de 100 indivíduos desenvolvendo em 100 gerações, tem-se que gerar 10.000 classificadores para finalizar a otimização. Nota-se que, dependendo do classificador utilizado a busca pode se tornar proibitiva. Isto costuma ocorrer com a utilização do classificador Máquinas de Vetores Suporte, que resolve uma otimização quadrática com restrição para gerar o hiperplano separador. Com um número grande de instâncias o custo para obter o resultado desta otimização é bastante alto.

Fica então, evidenciado, a necessidade de se buscar estratégias específicas para a diminuição do custo computacional da seleção de atributos. Uma solução de interesse quando se tem acesso a ambientes com arquiteturas paralelas é a utilização de algoritmos evolucionistas paralelos que já mostraram eficiência em termos de desempenho computacional sem perda na qualidade da resposta obtida.

A opção a ser desenvolvida aqui trata de um modelo de gerenciamento de aproximação onde uma parcela da população será avaliada e outra aproximada por técnicas com custo computacional inferior ao custo do problema original. A expectativa é que o custo computacional diminua proporcionalmente ao tamanho da parcela da população aproximada. Ao contrário do modelo em paralelo, deve-se avaliar a qualidade de resposta obtida com o método de aproximação, que pode gerar resultados de pouca acurácia devido à parcela que foi aproximada, a inadequação do método de aproximação ao problema ou a um gerenciamento da aproximação ineficiente.

Para a construção de um método de avaliação aproximada que funcione em conjunto com um algoritmo evolucionista devem-se definir os seguintes elementos:

- o gerenciador, que será acoplado ao algoritmo evolucionista e controlará todo o processo de aproximação, a saber:
  - seleção dos indivíduos a serem avaliados e aproximados;
  - atualização da base de dados utilizada para o cálculo da aproximação.
- especificação da base de dados referência para a aproximação;
- definição da técnica de aproximação.

Acrescenta-se ainda a função objetivo que irá definir a aptidão dos indivíduos e, no caso do problema de seleção de características, deve-se ainda determinar o classificador ou classificadores que serão utilizados para definir a aptidão do subconjunto de atributos ativos representado pelos indivíduos da população. Descreve-se, a seguir, cada um destes itens.

## 5.1. Algoritmo Genético para Seleção de Atributos

### 5.1.1 Função Objetivo:

Em relação ao AG, foi visto que a codificação binária dos atributos em alelos é a forma mais usual de representação e será utilizada. A função objetivo que indicará diretamente a aptidão dos indivíduos será definida como a porcentagem de acertos do classificador em questão para o conjunto de testes, escrita na forma:

$$f(x) = C_{trn(x)}(tst(x))$$

onde  $x$  é o indivíduo avaliado,  $tst(x)$  o conjunto que contém as instâncias para teste mapeadas de acordo com os atributos ativos de  $x$  e  $C_{trn(x)}$  a porcentagem de acerto no conjunto de testes do classificador gerado com um conjunto de treinamento considerando somente os atributos ativos de  $x$ . Esta função deve ser maximizada e seu valor fica na faixa  $[0,1]$ .

Optou-se por esta medida por ser uma função bastante simples e não levar em conta o número de atributos do indivíduo que gerou o classificador. A idéia é ter uma medida mais direta da qualidade do classificador gerado para avaliar melhor a qualidade dos métodos de aproximação utilizados.

Em experimentos adicionais adotou-se também, para efeito de comparação, uma função objetivo que privilegie indivíduos que tem menos atributos ativos. Geralmente este tipo de função é construído por meio da adição de um termo de penalização à função original, gerando:

$$f(x)_p = C_{trn(x)}(tst(x)) + \varepsilon \cdot (1 - nc(x))$$

onde  $\varepsilon$  é a constante de penalização e  $nc(x)$  a porcentagem das características ativas do indivíduo  $x$ . Apesar desta combinação ser bastante utilizada é conhecida a dificuldade de se determinar valores adequados para a constante de penalização. Técnicas adaptativas tendem a minorar tal dificuldade. Porém, optou-se por construir uma função de penalização não-linear que não necessita da constante de penalização dada por:

$$f(x)_{pnl} = (1 + C_{trn(x)}(tst(x))) \cdot (1 + (1 - nc(x))) = (1 + C_{trn(x)}(tst(x))) \cdot (2 + nc(x))$$

A adição das unidades visa prevenir multiplicações por zero, anulando a parcela não nula da função e que o produto reflita a maximização desejada.

Em relação aos parâmetros adotados para o AG, buscou-se adotar valores padrões que apresentassem uma busca razoável para esta classe de problema, evitando a otimização de tais parâmetros para cada banco de dados, o que não traria acréscimo à avaliação do modelo de aproximação. Tais valores foram fixados depois de alguns testes com os bancos utilizados. Exceção se faz para a estratégia de geração da população inicial, geralmente randômica. Utiliza-se, aqui, uma heurística para esta geração, específica para o problema de seleção de características, descrito a seguir.

### **5.1.2 População Inicial - Heurística:**

A expectativa na resolução de um problema de seleção de características é que a solução ótima apresente um número reduzido de características de forma a facilitar o entendimento dos atributos relevantes no contexto do problema tratado. Porém esta não é uma regra. Cada problema tem suas dependências específicas dos atributos para obter um resultado adequado na discriminação gerada pelo classificador. Desta forma, inicializar a população com poucos atributos pode dificultar a obtenção de soluções com muitos atributos devido a falta de diversidade de atributos ativos. O mesmo ocorre no caso contrário, onde a população inicial é construída com muitos atributos ativos. A solução randômica de distribuição uniforme parece mais sensata com metade dos atributos ativos e metade inativa, na média.

Testes realizados previamente indicaram que mesmo neste modelo randômico de geração inicial o aumento ou diminuição de atributos ativos se dá de forma lenta, caso esta seja a tendência da solução ótima para o problema. Propõe-se, então, uma heurística para geração inicial que apresente atributos ativos em três faixas, a saber:

- nível baixo de atributos ativos;
- nível médio de atributos ativos;
- nível alto de atributos ativos.

Assim, representa-se a diversidade em três níveis distintos de atributos ativos, ficando mais simples o encaminhamento da busca para a faixa ideal do problema tratado. Definem-se os níveis com os seguintes valores:

- nível baixo: 20% de atributos ativos;
- nível médio: 50% de atributos ativos;
- nível alto: 80% de atributos ativos.

O procedimento para a heurística de geração da população inicial é dado abaixo:

```

início
  nivelativos = 3;
  porc_at_ativos(1) = 0,2;
  porc_at_ativos(1) = 0,5;
  porc_at_ativos(1) = 0,8;
  para i de 1 ate natativos faça
    para j = 1, npop/nativos
      pop(j+i*nativos) = gera_rand_cromossomo(porc_at_ativos(j));
    fim para;
  fim para;
fim.

```

com `gera_rand_cromossomo(p)` sendo a função de geração randômica dos atributos com  $p\%$  ativos. A seguir apresenta-se, justifica-se e caracterizam-se os classificadores que serão utilizados.

## 5.2. Classificadores

Classificadores para o problema de discriminação tendem a ter comportamentos diferenciados dependendo de diversos aspectos relativos a sua aplicação. Números de instâncias, quantidade de atributos, tipo e magnitude dos atributos, número de classes, são alguns dos fatores que influenciam diretamente no desempenho dos classificadores tanto em termos de custo computacional quanto em relação a qualidade da resposta obtida.

Aplicando um modelo de aproximação em conjunto com o processo de seleção de características, espera-se que a sensibilidade dos classificadores em relação ao banco de dados utilizado seja ainda mais crucial para o nível de resposta obtida, pois soma-se a esta sensibilidade o nível de ruído ou erro presente nos indivíduos aproximados que também irão direcionar a busca pelo otimizador evolucionista.

Desta forma, visando avaliar a robustez do método de aproximação em relação ao classificador adotado, optou-se por trabalhar com três classificadores diferentes. A escolha dos classificadores foi feita de forma que os adotados tivessem propriedades e

estratégias de aplicação as mais diferenciadas possíveis. Descreve-se a seguir os classificadores adotados tentando ressaltar, principalmente, suas vantagens e desvantagens em seu uso.

### **5.2.1 Máquinas de Vetores Suporte (SVM)**

O classificador denominado Máquinas de Vetores Suporte (SVM), foi introduzida por Vapnik em 1992, formulada com todas as demonstrações matemáticas em seu livro (VAPNIK, 1995) e descrita em outro livro de sua autoria (VAPNIK, 1998) com maiores detalhes.

SVM é uma técnica de aprendizado de máquina, fundamentada nos princípios indutivos da minimização do risco estrutural. Estes princípios são provenientes da teoria do aprendizado estatístico, a qual está baseada no fato de que o erro da técnica de aprendizagem junto aos dados de validação (erro de generalização) é limitado pelo erro de treinamento mais um termo que depende da dimensão VC (dimensão Vapnik e Chervonenkis). A dimensão VC é uma medida da capacidade ou força de expressão de uma família de funções classificadoras obtidas por meio de um algoritmo de aprendizagem (VAPNIK e CHERVONENKIS, 1971).

De forma geral, o classificador SVM implementa um mapeamento não-linear (executado por um produto interno kernel escolhido a priori) dos dados de entrada para um espaço característico de alta-dimensão, em que um hiperplano ótimo é construído para separar os dados linearmente em duas classes. Quando os dados de treinamento são separáveis, o hiperplano ótimo no espaço característico é aquele que apresenta a máxima margem de separação. Para dados de treinamento em que as amostras das duas classes apresentam superposição (dados não separáveis), uma generalização deste conceito é utilizada. A técnica de aprendizado de máquina para determinação deste hiperplano ou superfície de decisão é denominada SVM, sendo que os dados de treinamento que se encontram à distância mínima do hiperplano são chamados vetores-suporte.

Principais vantagens:

- Generaliza bem mesmo treinado com reduzida quantidade de instâncias
- Pouco sensível a distribuição de probabilidade dos dados

Principais desvantagens:

- Dificuldade na escolha do kernel mais adequado

- Gasto computacional elevado, principalmente em bases com muitas instâncias

### 5.2.2 Vizinhos mais próximos (KNN)

O KNN classifica um dado elemento de acordo com os vizinhos mais próximos, onde  $k$  é o número de vizinhos. O algoritmo calcula a distância do elemento dado para cada elemento da base de treinamento, ordena os elementos da base de treinamento do mais próximo ao de maior distância. Dos elementos ordenados seleciona apenas os  $k$  primeiros, que servem de parâmetro para a regra de classificação. Utilizando por exemplo,  $k = 1$ , é selecionado apenas o elemento do treinamento mais próximo da instância que se pretende classificar. Já se for utilizado  $k=3$ , serão utilizados os três elementos mais próximos da instância e a determinação da classe é baseada nas classes dos três elementos determinados.

Principais vantagens:

- Simples de ser implementado
- Eficiente para conjunto de treinamento grande

Principais desvantagens:

- Necessidade de determinação quantidade de vizinhos a considerar
- Dificuldade na definição da métrica de distância a ser usada.
- Cômputo da distância de cada amostra em relação a todas do conjunto de treinamento.

### 5.2.3 Algoritmo das K-médias (K-means)

K-means é um algoritmo para classificar ou se agrupar objetos baseado em atributos, numa  $C$  quantidade de grupos. A idéia é fornecer uma classificação de informações de acordo com os próprios dados, baseada em análises e comparações entre os seus valores numéricos, sem considerar a pré-classificação existente. Por causa desta característica, o K-means é considerado como um algoritmo não supervisionado.

O passo básico de agrupar do K-means é simples. No princípio é determinado o número de agrupamentos  $C$  e o centróide ou centro destes agrupamentos. Pode-se considerar qualquer objeto como o centróide inicial. A seguir define-se a qual grupo cada uma das instâncias deve pertencer de acordo com a distância mínima desta em relação aos centróides considerados. Procede-se a atualização dos centróides, repetindo este processo até que nenhum dado seja movido do grupo.

Principais vantagens:

- Eficaz em grandes conjuntos de dados
- Normalmente converge com poucas iterações

Principais desvantagens:

- A quantidade de grupos,  $C$  deve ser determinada previamente.
- Uso de muita memória quando a quantidade de atributos é grande.
- Sensível a condição inicial.

### **5.3. Estrutura de Gerenciamento do Modelo de Aproximação**

O gerenciamento do modelo de aproximação é de vital importância para o bom desempenho das técnicas de aproximação. O controle dos elementos a serem simulados e avaliados deve trazer o balanço adequado visando um ganho no desempenho computacional com perda mínima no nível de resposta obtida. Descreve-se a seguir as características do gerenciamento adotado:

#### **5.4.1 Seleção dos indivíduos a Serem Avaliados e Aproximados**

Dada a população vigente em uma geração do AG, deve-se enviar os indivíduos para o gerenciador de forma que seja determinado aqueles que devem ser avaliados e os que devem ser aproximados baseado em uma porcentagem de avaliação ( $pav$ ) da população, que é um dos parâmetros de entrada do modelo. Optou-se por adotar um modelo determinístico na definição dos indivíduos a serem avaliados visando escolher os indivíduos que apresentem maior potencial para serem submetidos a avaliação ou simulação. Assim, toda a população deve ser inicialmente aproximada para que a aptidão da aproximação defina os indivíduos que apresentaram melhores resultados em relação ao modelo de aproximação. A porcentagem  $pav$  dos melhores indivíduos aproximados é enviada para simulação. Feita a simulação da porcentagem  $pav$  seu resultado substitui os resultados obtidos por meio da aproximação e a aptidão de todos os indivíduos da população é enviada para que o AG aplique os operadores genéticos de forma habitual.

Esta estratégia de seleção determinística gera um custo adicional de aproximação dos indivíduos que serão avaliados. Se o método de aproximação tiver um custo computacional bem menor do que o custo da simulação esta estratégia traz o benefício de uma escolha que enfoca a qualidade dos indivíduos da população sem grande comprometimento quanto ao desempenho computacional. Logicamente, se o

nível de erro no processo de aproximação for alto, a seleção pode ser equivocada, não compensando o custo computacional adicional.

Uma outra estratégia de seleção também foi implementada para comparação com os resultados obtidos pelo modelo determinístico. Trata-se de um modelo randômico onde, dada a população vigente, o gerenciador determina previamente a porcentagem *pav* da população que deve ser avaliada e envia os indivíduos para o simulador. Os demais são enviados para serem avaliados pelo modelo de aproximação. O gerenciador concatena as aptidões vindas da simulação e da aproximação e envia para o AG que aplicará os procedimentos genéticos para completar a geração. Os resultados deste modelo randômico serão apresentados em experimentos adicionais.

#### **5.4.2 Atualização da Base de Dados**

O modelo de aproximação utilizado irá determinar a aptidão dos indivíduos aproximados baseado nos dados contidos em uma base de dados construída para este fim, geralmente com indivíduos que foram avaliados/aproximados em gerações anteriores. Esta base geralmente possui tamanho constante e o gerenciador é responsável por sua atualização visando acompanhar o estágio em que se encontra a busca efetuada pelo AG. No modelo implementado a atualização é feita em cada geração de uma forma determinística. Finda a etapa de aproximação/avaliação pelo gerenciador, atualiza-se a base de dados utilizando-se somente os indivíduos avaliados. A atualização é feita por meio da inserção dos indivíduos avaliados na geração corrente que tenham aptidão melhor do que os piores indivíduos da base de dados. Caso os indivíduos avaliados não sejam melhores do que nenhum da base de dados, a mesma não será modificada. Maiores detalhes da base de dados para aproximação serão apresentados no item a seguir.

#### **5.4. Especificação, Construção e Caracterização da Base de Dados (População Auxiliar)**

Uma das premissas de algoritmos evolucionistas é a existência de uma população que deve evoluir com a aplicação de operadores genéticos em um processo geracional. Tal população pode ser pensada como uma base de dados representativa do estágio atual do processo evolutivo. Métodos de aproximação precisam de uma base de dados para efetuar a aproximação. Pode-se dizer, então, que métodos de aproximação combinados com algoritmos evolucionistas devem manipular duas bases de dados ou populações distintas ou unificar ambas em somente uma. Estudos prévios (FONSECA,

2009) indicaram que a unificação não é interessante porque a base de dados para o método de aproximação funciona melhor com tamanhos superiores à população do algoritmo evolucionista. Visto que a unificação não é interessante em termos de desempenho, as bases serão unificadas pelo menos no que se refere a nomenclatura, com a base de dados da aproximação sendo denominada população auxiliar. Ou seja, tem-se agora a população do AG e a população auxiliar do modelo de aproximação.

Inicialmente, deve-se determinar o tamanho da população auxiliar que permanecerá constante. De acordo com Fonseca (2009), uma escolha que mostrou-se adequada é a população auxiliar duas vezes o tamanho da população utilizada no algoritmo evolucionista. Testes preliminares nos problemas tratados aqui mostraram que esta proporção também foi adequada e, portanto, adotada.

Visando diminuir o nível de ruído no modelo de aproximação, a população auxiliar será formada somente por indivíduos que passaram pela simulação. Desta forma, nas duas primeiras gerações do AG todos os indivíduos são avaliados e aproveitados para a construção da população auxiliar inicial. A partir daí todo o processo de atualização é controlado pelo gerenciador conforme descrito anteriormente.

A evolução da população inicial no decorrer das gerações apresenta características interessantes em relação a população do AG. A política de atualização adotada leva a um desenvolvimento mais suave das aptidões representadas por não conter indivíduos aproximados e por substituir somente os piores indivíduos. Gera-se também um elitismo natural onde garante-se que o melhor indivíduo avaliado estará representado na população auxiliar. Desta forma o acompanhamento da aptidão dos indivíduos da população auxiliar pode até mesmo substituir o tradicional acompanhamento dos indivíduos da população do AG que, por incluir indivíduos aproximados e ser geracional apresenta uma evolução menos suave.

## **5.5. Modelo de Aproximação**

Diversos foram os modelos de aproximação descritos no capítulo anterior. É de maior interesse e facilidade quando se trabalha em conjunto com algoritmos evolucionistas que a abordagem de construção seja feita com foco no ajuste de dados. Isto se explica pela facilidade de montagem da base de dados (população auxiliar) quando se tem a população inerente dos algoritmos evolucionistas, sendo este o caminho adotado com aproximação evolucionista.

Entre os modelos baseados em similaridade também é bastante natural e simples a aplicação de uma estratégia de herança de aptidão. Porém, optou-se aqui por utilizar dois outros modelos. O primeiro é baseado no conhecido vizinho mais próximo e o segundo é uma técnica original proposta aqui com enfoque específico nos alelos do cromossomo e que pretende-se que tenha um bom desempenho para o problema de seleção de características tratado neste trabalho. Os modelos, principalmente o segundo serão apresentados a seguir.

### 5.6.1 Aproximação pelos Vizinhos mais Próximos (VMP)

O modelo usado é similar ao descrito no capítulo anterior. Dado um indivíduo a ser aproximado calcula-se a distância dele para os indivíduos da população auxiliar. Pelas características de codificação do problema de seleção a medida escolhida para a determinação da distância é a medida de *Hamming*.

A distância de *Hamming* consiste no número de posições em que duas seqüências de bits de mesmo tamanho diferem. Exemplificando, dadas as seqüências  $x_1 = 11011101$  e  $x_2 = 11000101$ , a distância de *Hamming* entre  $x_1$  e  $x_2$ , denotada por  $Hamming(x_1, x_2)$ , é 2. Basta verificar que os bits de  $x_1$  e  $x_2$  diferem nas posições 4 e 5.

Assim, quanto mais atributos ativos e inativos dois indivíduos tiverem em comum, mais próximos eles estarão. A aptidão então é calculada pela média da aptidão dos indivíduos mais próximos da população auxiliar em relação ao indivíduo  $x$ , em questão:

$$apt_{apv}(x) = \frac{\left( \sum_{i=1}^k (apt(cl_{cr}(x_{popaux}))) \right)}{k}$$

onde  $cl_{cr}(x_{popaux})$  é a ordenação crescente dos indivíduos da população auxiliar que minimizam a distância de *Hamming* em relação ao indivíduo aproximado  $x$ ,  $k$  o número de vizinhos considerados e  $apt_{apv}(x)$  a aptidão aproximada pelos vizinhos mais próximos para o cromossomo  $x$ . Testes preliminares indicaram que a consideração de três indivíduos mais próximos da população auxiliar apresenta resultados satisfatórios.

### 5.6.2 Aproximação por Alelo (ou Atributo)

O método de aproximação proposto é construído para funcionar de forma adequada, ou seja, obter uma boa aproximação em problemas em que os cromossomos apresentam baixa correlação entre seus genes/alelos (baixa epistasia). A aplicação

satisfatória em problemas de seleção de característica baseia-se na baixa epistasia presente entre os genes/alelos que representam os atributos no cromossomo. Tanto o conceito de epistasia em cromossomo quanto à verificação da efetividade do modelo em relação a problemas sintéticos serão apresentados após a definição da estratégia de aproximação por alelo.

A aproximação por alelo é construída enfocando-se a importância de cada alelo do cromossomo na aptidão obtida em indivíduos de uma população. A primeira etapa é um pré-processamento da população para determinar a aptidão de cada alelo de acordo com a população vigente. No caso binário, deve-se gerar a aptidão para cada alelo caso o valor seja 0 e a aptidão para cada alelo caso seu valor seja 1. Suponha que esteja sendo calculado a aptidão para um determinado alelo  $a_i$  de um cromossomo. Deve-se varrer todos os indivíduos da população auxiliar verificando os que na posição do alelo  $a_i$  possuem valor 0, gerando o somatório:

$$ap0(a_i) = \frac{\sum_j apt(x_j)}{\#(a_i = 0)} \quad j = 1, \dots, npopaux$$

com  $ap0(a_i)$  sendo a aptidão do alelo  $a_i$  para o valor 0,  $\#(a_i=0)$  é a cardinalidade do conjunto que tem o alelo  $a_i$  igual a 0 e  $npopaux$  o tamanho da população auxiliar. O valor de  $ap0$  deve ser calculado para todos os alelos do cromossomo.

Em seguida faz-se o mesmo para os alelos  $a_i$  da população com valor 1:

$$ap1(a_i) = \frac{\sum_j apt(x_j)}{\#(a_i = 1)} \quad j = 1, \dots, npopaux$$

com  $ap1(a_i)$  sendo a aptidão do alelo  $a_i$  para o valor 1.

Finda esta etapa, tem-se dois vetores ( $ap0$  e  $ap1$ ) com tamanho igual ao número de alelos do cromossomo, que guardam a aptidão por alelo para o caso do alelo ser 0 ou 1, respectivamente.

No caso do modelo de aproximação, estes vetores são calculados com base na população auxiliar, que serve de base para o cálculo da aproximação.

A segunda etapa calcula efetivamente a aptidão aproximada de um indivíduo  $x$  com base nos vetores  $ap0$  e  $ap1$ , na seguinte forma:

$$apt_{apa}(x_i) = \frac{\sum_{\forall x_i=0}^i ap0(a_i) + \sum_{\forall x_i=1}^i ap1(a_i)}{L} \quad i = 1, \dots, L$$

sendo  $apt_{apa}(x)$  a aptidão aproximada por alelo para o cromossomo  $x$ . Como esta forma de definição da aptidão aproximada é construída com o enfoque na importância do alelo na determinação da aptidão do indivíduo, caso haja correlação entre alelos dificilmente esta correlação será capturada na construção dos vetores  $ap0$  e  $ap1$ . Desta forma, espera-se que a qualidade da aproximação nestes casos seja deteriorada.

No caso de aplicação em problemas de seleção de características sugere-se que este modelo de aproximação seja chamado de aproximação por atributo pois, neste caso, cada alelo representa um atributo das instâncias da base de dados. Apresenta-se, a seguir, um exemplo numérico para ilustrar o cálculo da aptidão por atributo.

Suponha que os indivíduos abaixo representem a população de cromossomos com suas devidas aptidões obtida por simulação:

- 1101 → 70%
- 1000 → 60%
- 0011 → 50%
- 0110 → 40%
- 0001 → 30%

**1ª etapa:** Cálculo da aptidão para cada atributo:

1º. Atributo  $ap0(1) = (50 + 40 + 30)/3 = 40.00$

$ap1(1) = (70 + 60)/2 = 65.00$

2º. Atributo  $ap0(2) = (60 + 50 + 30)/3 = 46.67$

$ap1(2) = (70 + 40)/2 = 55.00$

3º. Atributo  $ap0(3) = (70 + 60 + 30)/3 = 53.33$

$ap1(3) = (50 + 40)/2 = 45.00$

4º. Atributo  $ap0(4) = (60 + 40)/2 = 50.00$

$ap1(4) = (70 + 50 + 30)/3 = 50.00$

Assim  $ap0 = (40.00 \ 46.67 \ 53.33 \ 50.00)$

e  $ap1 = (65.00 \ 55.00 \ 45.00 \ 50.00)$

**2ª etapa:** Indivíduo a ser avaliado por aproximação  $\rightarrow x = 1001$

Aptidão calculada para o 1º Atributo  $ap1(1) \rightarrow 65.00$

Aptidão calculada para o 2º Atributo  $ap0(2) \rightarrow 46.67$

Aptidão calculada para o 3º Atributo  $ap0(3) \rightarrow 53.33$

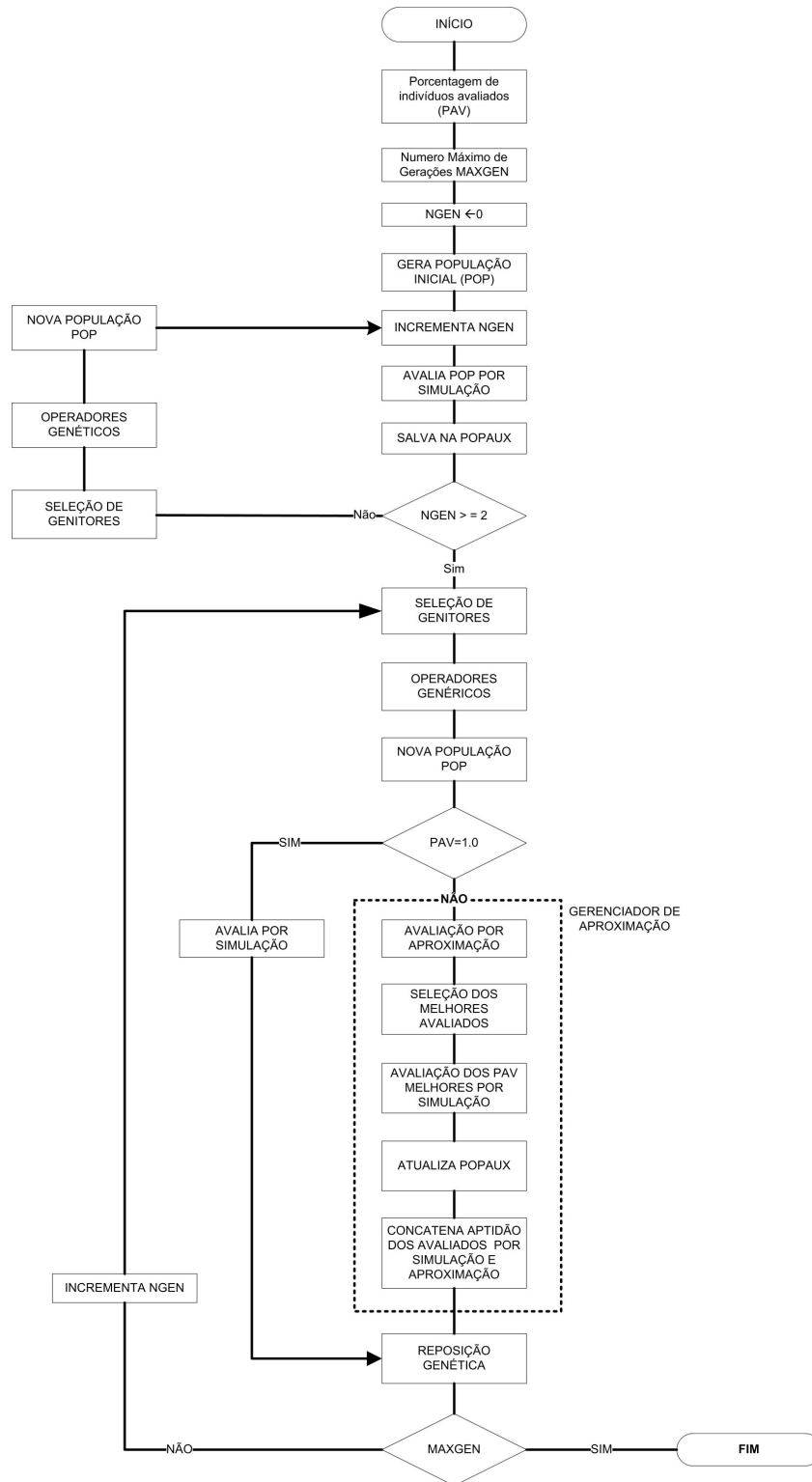
Aptidão calculada para o 4º Atributo  $ap1(4) \rightarrow 50.00$

Assim, o cálculo da aptidão aproximada de  $x$  será:

$(65.00 + 46.67 + 53.33 + 50.00)/4 = 53.75 \rightarrow$  Aptidão aproximada por atributo

## 5.6. Algoritmo Implementado

O fluxograma abaixo representa a solução implementada.



A seguir apresenta-se um estudo desta estratégia de aproximação aplicada em problemas sintéticos com níveis diferentes de epistasia.

### 5.7. Aptidão por Alelo (Atributo) e Epistasia

O fenômeno de epistasia (KAUFFMAN, 1993) é conhecido como a interação entre genes ou alelos de um cromossomo e, geralmente, tende a tornar o problema de otimização mais complexo e de difícil solução.

O método de avaliação aproximada por alelo ou atributo, quando se trata de aplicação em problemas de classificação, tende a ser mais sensível em problemas que possuem níveis maiores de epistasia. Tal avaliação é de simples constatação, bastando analisar a forma de definição da aptidão usada no método, onde a composição da aptidão do cromossomo é feita focando-se em um alelo de cada vez e combinando estes resultados individuais.

Porém, espera-se que tal característica deste método de aproximação proposto não venha a comprometer sua utilização em problemas de seleção de atributos utilizando AG. Isto porque considera-se que cada atributo funciona de uma forma razoavelmente independente em relação aos outros na definição da classificação da instância.

De qualquer forma, é interessante testar o desempenho do modelo em problemas binários, de fácil entendimento e controle para se ter uma real idéia do comportamento do modelo diante de diferentes níveis de epistasia. Dada uma cadeia binária  $x = (x_1 x_2 \dots x_L)$  determinam-se as funções:

- **MaxUns** ou máximo de uns, onde busca-se a maximização de uns em uma cadeia binária. Este problema tem baixo nível de epistasia. É definido na forma:

$$f_{uns}(x) = \sum_{i=1}^L x_i$$

- a outra função é conhecida como função **Royal Road**(FORREST e MITCHELL, 1993) e apresenta um nível médio de epistasia. Esta função envolve a definição de um grupo pré-determinado de esquemas  $S = (s_1, s_2, \dots)$ . Será utilizada sua versão que considera blocos de 8 bits ou alelos que definem estes esquemas com ordem de tamanho 8. Desta forma, a cadeia binária  $x$  de  $L$  bits deve ser múltipla de 8. Estes esquemas servirão de base para gerar o valor da função por meio da composição:

$$f_{rr} = \sum_{s \in S} c_s \sigma_s(x)$$

onde  $c_s$  é o valor designado para o esquema  $s$  sendo adotado  $c_s = \text{ordem}(s) = 8$ , conforme mostrado na Figura 5.2 para uma cadeia de 64 bits e  $\sigma_s$  é dado por:

$$\sigma_s = \begin{cases} 1 & \text{se } x \text{ é instância de } s \\ 0 & \text{caso contrário} \end{cases}$$

Ou seja,  $\sigma_s$  vale 1 se  $x$  é uma instância de  $s$  e 0 caso não seja.

```

S1 = 11111111 *****; C1 = 8
S2 = *****11111111 *****; C2 = 8
S3 = *****11111111 *****; C3 = 8
S4 = *****11111111 *****; C4 = 8
S5 = *****11111111 *****; C5 = 8
S6 = *****11111111 *****; C6 = 8
S7 = *****11111111 *****; C7 = 8
S8 = *****11111111; C8 = 8

```

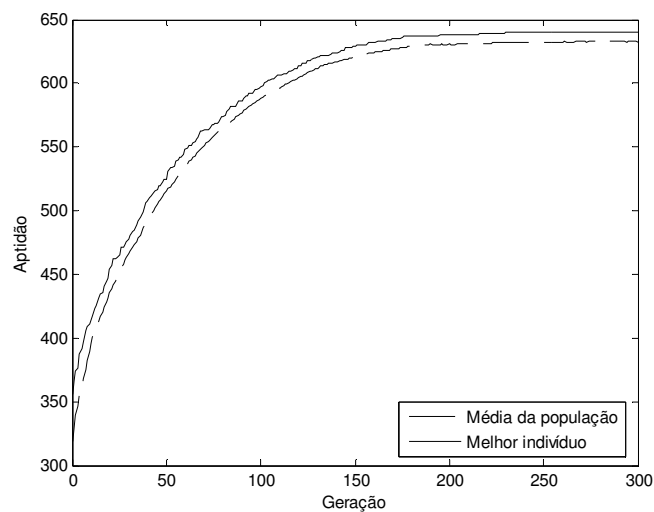
Figura 5.2 - Conjunto de esquemas da função **Royal Road**

Os testes serão obtidos usando um AG coma as seguintes características:

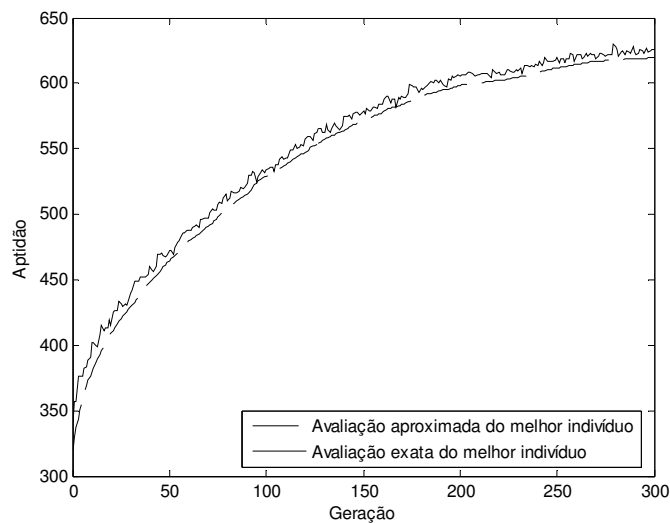
- tamanho da cadeia binária: 640
- tamanho da população: 500
- número de gerações: 300 (**MaxUns**) e 1000 (**Royal Road**)
- probabilidade de recombinação: 1,0
- probabilidade de mutação: 0,003 por bit

Os resultados representam uma execução padrão do algoritmo. Não houve a preocupação de aplicação de métodos de validação por meio de médias entre diversas execuções porque queria-se apenas ter uma idéia do perfil do comportamento do modelo de aproximação entre estes dois casos com níveis diferenciados de epistasia. A aplicação segue um formato diferente dos modelos tradicionais de gerenciamento de aproximação. Neste caso não se tem base de dados (população auxiliar) e toda a população é aproximada baseada na aptidão correta dos indivíduos, ou seja, não existe uma parcela da população que é efetivamente avaliada. A figura 5.3 mostra o resultado da função **MaxUns** avaliando-se todos os indivíduos, isto é, sem aproximação. Consegue-se obter, neste caso, o resultado ótimo ou aptidão igual a 640. Na figura 5.4 apresenta-se o resultado da aproximação em conjunto com o valor que o melhor

indivíduo por geração obteria caso fosse avaliado. Na última geração obteve-se como resultado o melhor valor de aptidão aproximada de 619,7785 que, se fosse avaliado teria a aptidão de 626. O erro entre o valor aproximado e o valor avaliado fica na faixa de 1%. O erro em relação ao caso onde avaliam-se todos os indivíduos é da ordem de 2%. Nota-se que o resultado da aproximação é sempre menor do que o valor avaliado e que, nem sempre, a melhor aproximação obtida indica que se terá melhor avaliação. No exemplo apresentado, o melhor indivíduo avaliado obteve aptidão de 630, melhor que o resultado final obtido.



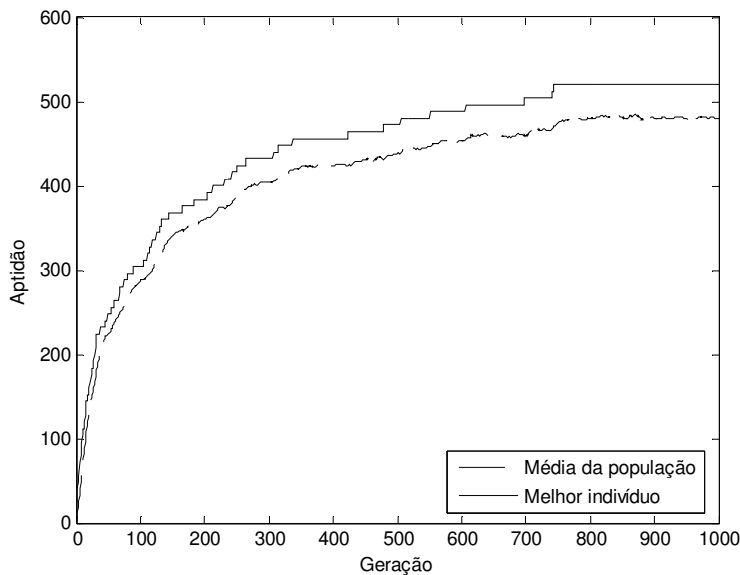
**Figura 5.3 - Função MaxUns: população avaliada**



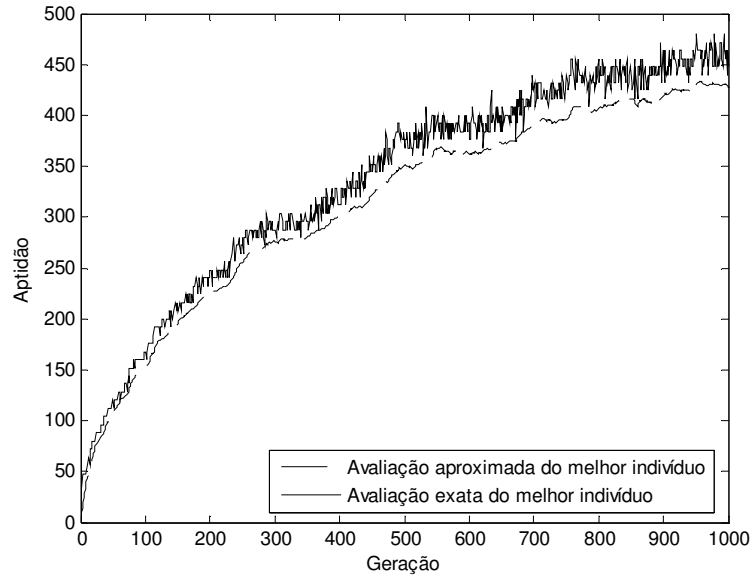
**Figura 5.4 - Função MaxUns: população aproximada**

A figura 5.5 mostra o resultado para a população avaliada para a função **Royal Road**. O resultado ótimo (640) não foi alcançado nas 1000 gerações, sendo 520 o ótimo obtido nesta execução. A figura 5.6 indica o resultado para o modelo de aproximação. O melhor indivíduo aproximado obteve a aptidão de 427,3294 que, avaliado geraria a aptidão de 448, indicando uma diferença em torno de 4,5%. O erro entre o resultado avaliado e o aproximado foi em torno de 14%. O melhor indivíduo do modelo aproximado, caso fosse avaliado, apareceu em gerações intermediárias com aptidão de 480.

Fazendo-se uma análise final, pode-se dizer que os resultados obtidos estão de acordo com as expectativas descritas sobre como o modelo aproximado deveria funcionar em presença de níveis diferentes de epistasia. O resultado para a função **MaxUns** apresentou níveis de erro tanto na qualidade da aproximação quanto em relação ao resultado final em relação ao modelo avaliado bem menores do que os da função **Royal Road**, que apresenta níveis maiores de epistasia. Acredita-se, desta forma, que aplicado em uma estratégia de seleção de características apresentará resultados competitivos tanto em termos de qualidade da resposta quanto em relação ao custo computacional. Deve-se ainda, considerar que no modelo de gerenciamento de aproximação uma parcela da população é avaliada, melhorando a precisão do resultado.



**Figura 5.5** - Função **Royal Road**: população avaliada



**Figura 5.6** - Função **Royal Road**: população aproximada

No capítulo a seguir, apresentam-se experimentos numéricos visando avaliar o desempenho e robustez do modelo de seleção de características com avaliação aproximada proposto, aplicados em bancos de dados com uma grande variedade de números de atributos e instâncias.

## CAPITULO 6 Experimentos Numéricos

Apresentam-se, neste capítulo, os testes realizados com a estratégia proposta para seleção de características utilizando algoritmos genéticos associado a um modelo de aproximação. Objetiva-se avaliar o potencial do modelo aplicando-o em diversas bases de dados com perfis bastante diferenciados. Enfoca-se a utilização em bases de expressão genética em *microarray* que, geralmente, são compostas por poucas instâncias e um número elevado de características.

Em função de testes preliminares nas bases de dados que serão utilizadas, foram fixados alguns parâmetros do AG geracional com codificação binária, a saber:

- Tamanho da população: 30 indivíduos
- Número máximo de gerações: 100
- Probabilidade de Recombinação: 80% (um ponto)
- Probabilidade de Mutação: 10%
- Seleção: SUS
- Aplicação de Elitismo

Também para os classificadores, alguns parâmetros foram fixados:

- Classificador KNN, números de vizinhos igual a 3
- Classificador K-means, número de centróides igual a 2
- Classificador SVM, kernel Gaussiano

Deve-se ressaltar que não se tinha o objetivo de otimizar os parâmetros para cada base de dados utilizada mas, definir parâmetros que apresentassem um nível adequado de desempenho.

### 6.1. Recursos Computacionais

Os testes realizados foram feitos em um computador com processador Pentium Core 2 Duo, velocidade de 2.1 Ghz com 3.0 Gigabytes de memória RAM. O software

utilizado na programação foi o MATLAB 7.0, rodando no sistema operacional Windows XP.

## 6.2. Bases de Dados

As bases utilizadas para testes estão na tabela abaixo, todas com duas classes.

NOME	NATUREZA DOS DADOS	ATRIBUTOS	INSTÂNCIAS	ORIGEM
Breast	Reais	12625	24	Ludwig Institute for Cancer Research
Colon	Reais	2001	62	
Leukemia	Inteiros	7130	72	
Prostate	Reais	12601	102	
Ionosphere	Reais	35	351	Site UCI Machine Learning Repository <a href="http://archive.ics.uci.edu/ml/datasets.html">http://archive.ics.uci.edu/ml/datasets.html</a>
Mushroom	Binários	99	5644	
Sonar	Reais	61	208	
Synthetic	Reais	61	600	

**Tabela 6.1** - Bases para testes

## 6.3. Experimentos

Os resultados apresentados são obtidos através da média de um procedimento de validação cruzada com três partes. Os experimentos são realizados sem aproximação, e com as estratégias de aproximação VMP e por ATRIBUTOS. Quando considera-se a aproximação, são utilizados os percentuais de 60%, 40% e 20% para a quantidade de indivíduos que serão avaliados. Inicialmente, assume-se como aptidão a porcentagem de acertos do classificador. O gerenciamento determinístico adotado permite o controle do erro de aproximação dos indivíduos avaliados, que também será apresentado. A seguir, seguem-se os resultados das primeiras bases avaliadas.

### 6.3.1 Bases de Dados Sonar, Breast e Ionosphere

Com estas bases foram executadas todas as variações de aproximações previstas, descritas anteriormente. São bases com quantidade expressiva de instâncias e razoável quantidade de características. Ressalta-se, em particular, a base Breast, por se tratar de uma aplicação em Biologia formada por base de expressão genética em *microarray*, geralmente caracterizada por ter poucas instâncias e elevado número de atributos.

a) Base Sonar

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Aptidão (%)	Erro médio da aproximação (%)
KNN	SEM	100	15.51	55.0	90.0	-----
	VMP	60	11.46	50.0	96.0	15.19
		40	7.47	45.0	94.0	13.83
		20	4.45	45.0	93.0	12.45
	ATRIBUTOS	60	10.70	52.0	96.0	9.66
		40	7.99	38.0	94.0	9.53
20		3.84	42.0	94.0	8.49	
K-means	SEM	100	30.24	33.0	71.0	-----
	VMP	60	19.78	40.0	80.0	30.50
		40	14.56	48.0	78.0	30.08
		20	8.69	43.0	77.0	30.49
	ATRIBUTOS	60	18.31	22.0	77.0	20.98
		40	13.93	50.0	78.0	20.26
20		6.83	27.0	81.0	19.88	
SVM	SEM	100	5510.19	35.0	94.0	-----
	VMP	60	3337.17	47.0	97.0	8.86
		40	2261.22	33.0	94.0	7.67
		20	1192.19	50.0	97.0	10.14
	ATRIBUTOS	60	3370.44	48.0	97.0	6.54
		40	2271.69	45.0	99.0	7.37
20		1192.07	53.0	97.0	5.87	

**Tabela 6.2-** Base de Dados Sonar 600 instâncias e 60 atributos

Na tabela 6.2 observa-se, de maneira geral, que para a maioria dos testes realizados, a quantidade de atributos obtidos ficou na faixa de 50%. O classificador de pior desempenho foi o K-means, produzindo maior erro de aproximação e menores índices de acerto. O melhor desempenho coube ao classificador SVM, porém a um custo computacional elevadíssimo se comparado aos demais (tempo em segundos). Nota-se, porém, que não houve mudança no padrão de comportamento dos classificadores, independente do tipo de aproximação e da quantidade de indivíduos avaliados utilizada.

Os modelos de aproximação adotados mostraram-se eficientes, com pequena vantagem para o modelo de ATRIBUTOS. Em termos de custo computacional, os resultados são altamente relevantes, com ganhos proporcionais a porcentagem de indivíduos aproximados. Deve-se ressaltar que, ao contrário do esperado, resultados com aproximação obtiveram melhor índice de acerto do que os casos sem aproximação, para todos os classificadores utilizados. Acredita-se que isto ocorra devido à forma de seleção dos indivíduos a serem avaliados implementada e sua combinação na população com os indivíduos somente aproximados, que geralmente apresentam aptidão com valores inferiores ao que seria obtido com a avaliação. Tal combinação gera uma nova dinâmica no comportamento do AG, permitindo alcançar tais níveis de resultados.

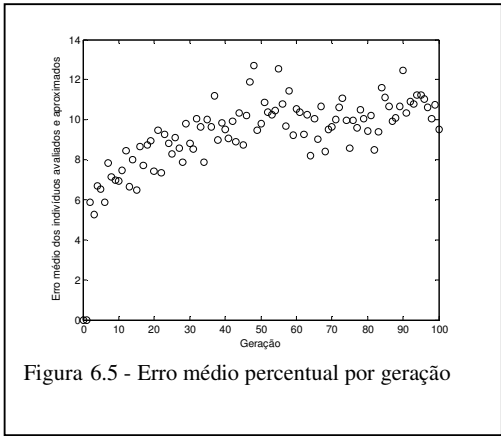
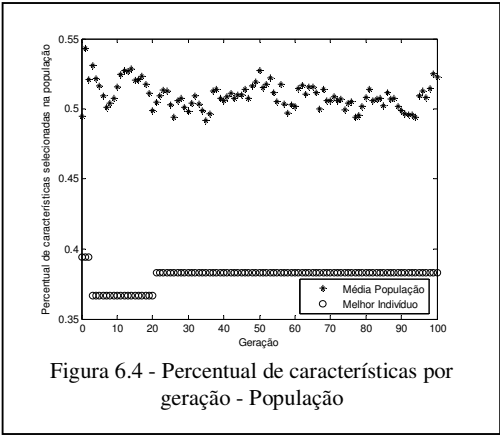
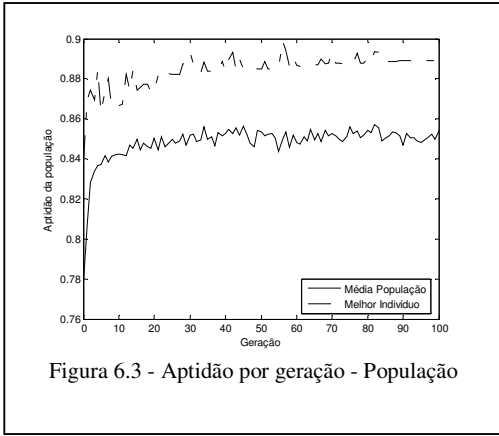
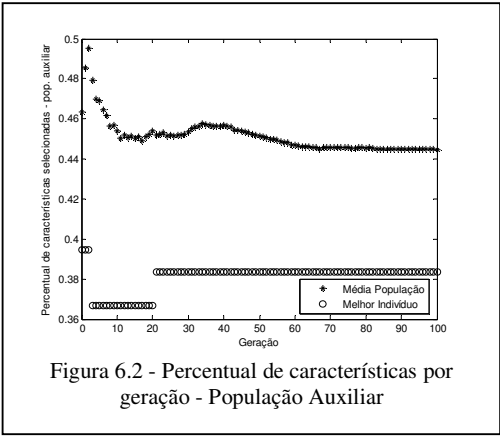
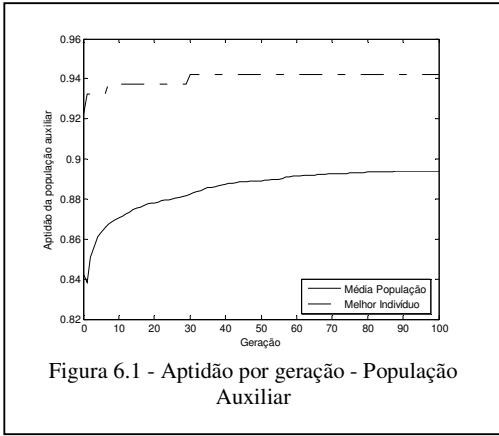
Em relação aos erros de aproximação, observa-se que para menores porcentagens a serem avaliadas o erro tende a ser menor. Justifica-se o fato, devido à maior similaridade entre os melhores indivíduos da população escolhidos para avaliação com a população auxiliar, que é construída somente com os melhores indivíduos avaliados durante todo o processo evolutivo. Desta forma, aumenta-se a qualidade da aproximação para estes melhores indivíduos por estarem mais bem representados na população auxiliar. Com o aumento da porcentagem de indivíduos a serem avaliados, elementos menos similares à população auxiliar também são alocados para avaliação, o que tende a aumentar o erro. Ressalta-se que, o erro é calculado somente para os indivíduos que foram aproximados e avaliados na geração.

Abaixo, os gráficos 6.1 a 6.5 obtidos com a combinação:

Classificador:.....K-means

Aproximação:.....ATRIBUTOS

Percentual Avaliado:.....40%



b) Base Breast

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Aptidão (%)	Erro médio da aproximação (%)	
KNN	SEM	100	79.05	20.0	88.0	-----	
	VMP	60	232.65	20.0	100.0	23.77	
		40	213.61	20.0	100.0	25.79	
		20	192.67	20.0	100.0	27.22	
	ATRIBUTOS	60	85.85	20.0	100.0	13.55	
		40	72.10	20.0	100.0	15.02	
		20	57.24	20.0	100.0	16.92	
	K-means	SEM	100	188.61	20.0	96.0	-----
		VMP	60	283.05	20.0	100.0	49.76
40			253.21	20.0	100.0	50.19	
20			211.54	20.0	100.0	51.23	
ATRIBUTOS		60	153.91	20.0	100.0	40.82	
		40	118.09	19.0	100.0	39.39	
		20	87.21	80.0	100.0	39.33	
SVM		SEM	100	609.80	20.0	83.0	-----
		VMP	60	685.69	20.0	100.0	43.26
	40		451.98	20.0	100.0	47.47	
	20		363.90	20.0	100.0	52.22	
	ATRIBUTOS	60	480.59	20.0	100.0	28.16	
		40	320.15	20.0	100.0	31.12	
		20	198.28	20.0	100.0	29.56	

**Tabela 6.3** Base de Dados Breast 24 instâncias e 12625 atributos

Na tabela 6.3 apresentada acima, observa-se que a qualidade dos resultados foi idêntica para todas as variações dos modelos com aproximação utilizadas, com acerto completo de todo conjunto de teste. Nos casos sem aproximação, novamente a qualidade apresentou-se um pouco pior. A maioria dos casos, com e sem aproximação, selecionou 20% dos atributos como melhor solução.

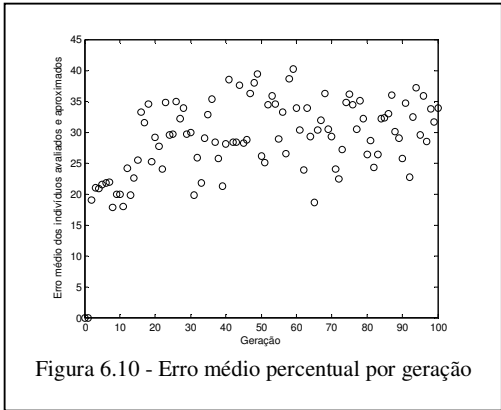
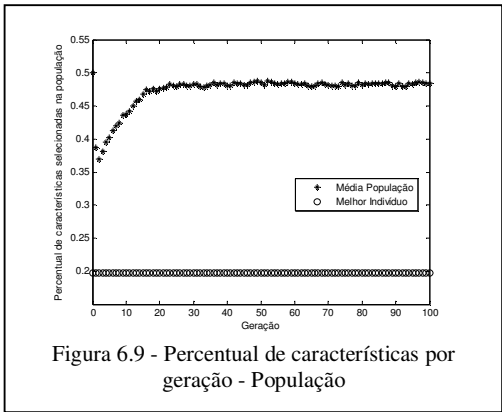
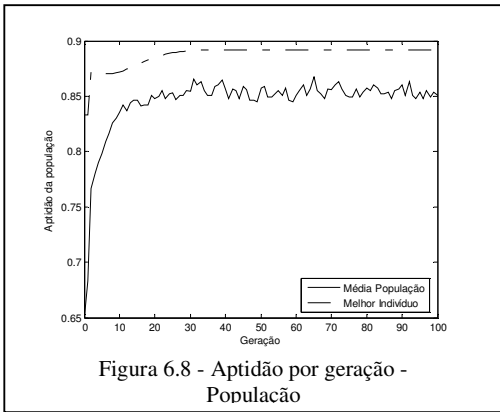
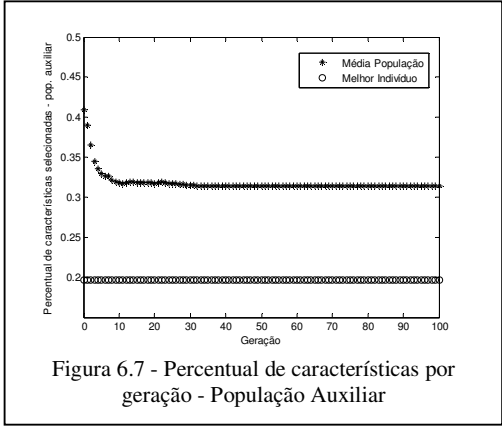
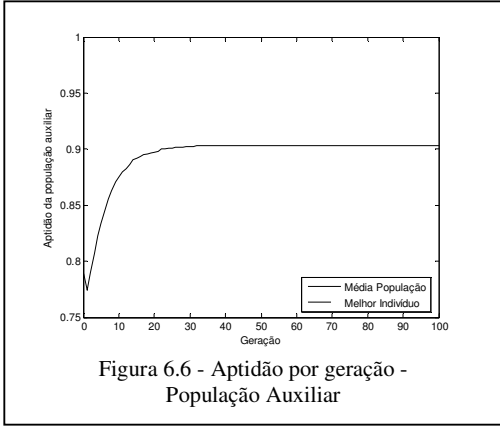
Os níveis de erro da aproximação por ATRIBUTO foram menores do que a aproximação VMP, bem como o custo computacional. Ressalta-se que, o custo computacional sem aproximação foi menor do que muitos casos dos modelos com aproximação, devido à base de dados ser composta por poucas instâncias com muitos atributos.

Abaixo, os gráficos 6.6 a 6.10 obtidos com a combinação:

Classificador:.....SVM

Aproximação:.....ATRIBUTOS

Percentual Avaliado:.....20%



c) Base Ionosphere

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Aptidão (%)	Erro médio da aproximação (%)
KNN	SEM	100	34.00	45.0	90.0	-----
	VMP	60	20.81	38.0	92.0	5.23
		40	16.71	50.0	91.0	5.28
		20	9.67	50.0	92.0	5.53
	ATRIBUTOS	60	26.24	45.0	92.0	4.67
		40	16.93	44.0	92.0	4.61
		20	9.03	42.0	93.0	4.23
K-means	SEM	100	27.79	22.0	79.0	-----
	VMP	60	18.41	53.0	84.0	15.84
		40	12.99	28.0	85.0	17.18
		20	7.54	41.0	81.0	13.26
	ATRIBUTOS	60	17.18	18.0	86.0	10.99
		40	11.68	24.0	84.0	9.23
		20	6.40	43.0	85.0	10.70
SVM	SEM	100	18709.61	44.0	96.0	-----
	VMP	60				
		40				
		20	4496.65	62.0	97.0	5.06
	ATRIBUTOS	60				
		40				
		20	4440.44	53.0	98.0	3.34

**Tabela 6.4** Base de Dados Ionosphere 351 instâncias e 35 atributos.

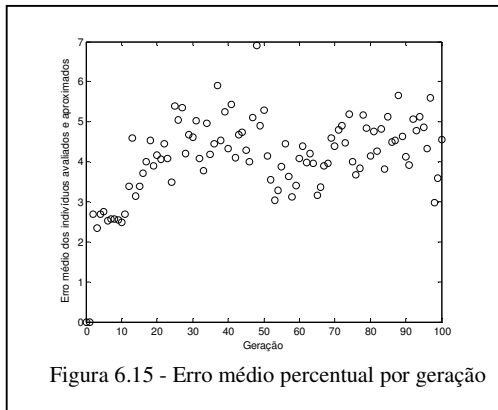
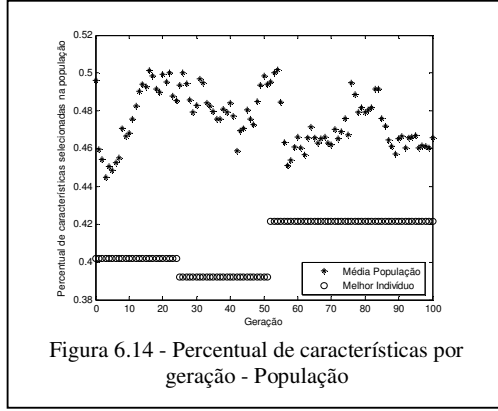
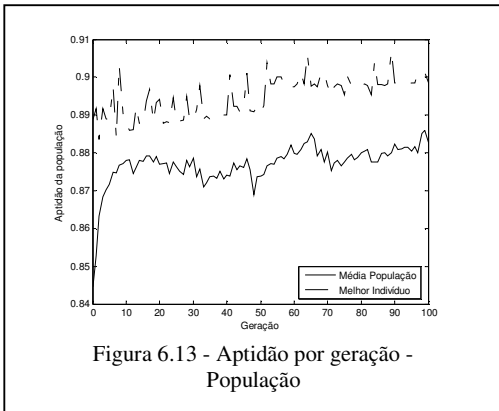
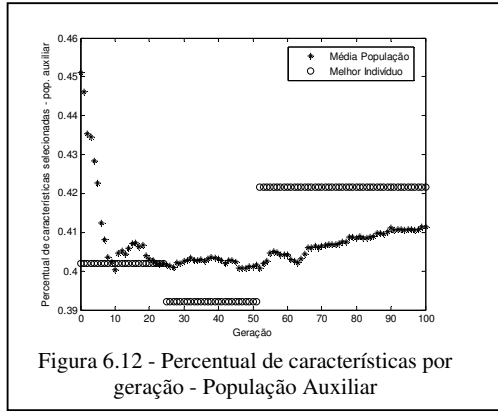
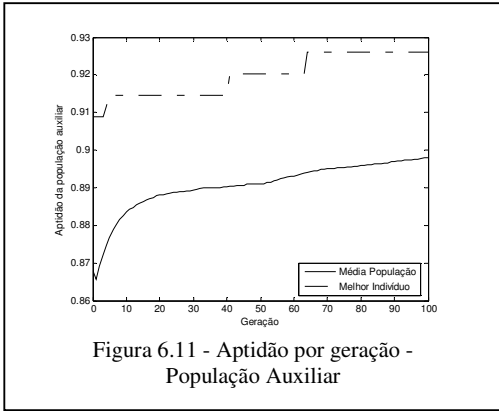
Na tabela 6.4, verifica-se um melhor desempenho do classificador SVM, porém com tempos de processamento elevadíssimos. Tão elevados que foram executados apenas os casos sem aproximação e as aproximações por ATRIBUTO e VMP com avaliação de 20% da população. Pode-se observar que para o modelo sem aproximação, considerando a validação cruzada de três partes, o tempo necessário para se executar este classificador foi em torno de 16 horas.

Novamente, ponderando os resultados obtidos e os tempos de execução, a técnica de aproximação por ATRIBUTOS se mostra mais eficiente em relação ao modelo VMP tanto em qualidade de resposta quanto em custo computacional.

O erro médio de aproximação foi mais baixo tanto para o KNN quanto para o SVM, apresentando valor mais elevado para o classificador K-means.

Abaixo, os gráficos 6.11 a 6.14 para a combinação que, no conjunto, apresentou um bom desempenho:

Classificador:..... KNN  
Aproximação:..... ATRIBUTOS  
Percentual avaliado:.....20%



### 6.3.2 Outras Bases de Dados

#### a) Base Synthetic

Esta base foi testada com os classificadores KNN e K-means, não sendo utilizado o classificador SVM, pois o tempo de processamento seria inviável. Abaixo, a tabela 6.5 mostra o resumo destes testes.

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Aptidão (%)	Erro médio da aproximação (%)	
KNN	SEM	100	103.56	41.0	95.0	-----	
	VMP	60	83.74	55.0	96.0	7.11	
		40	54.98	42.0	95.0	5.66	
		20	28.93	26.0	95.0	6.23	
	ATRIBUTOS	60	81.48	45.0	96.0	4.46	
		40	58.71	67.0	95.0	4.09	
		20	29.10	52.0	95.0	3.86	
	K-means	SEM	100	55.47	48.0	84.0	-----
		VMP	60	33.20	47.0	86.0	8.31
40			29.65	53.0	84.0	6.73	
20			14.61	55.0	85.0	6.01	
ATRIBUTOS		60	38.57	73.0	86.0	5.40	
		40	27.35	53.0	85.0	4.01	
		20	13.27	18.0	86.0	5.59	

**Tabela 6.5** - Base de Dados Synthetic 600 instâncias e 60 atributos

Analisando-se os resultados na tabela acima, as execuções com o classificador KNN se mostram mais eficientes e, novamente, a combinação com a aproximação por ATRIBUTOS apresenta um melhor desempenho. Em alguns casos, o tempo de execução é ligeiramente maior neste modelo sendo este fato devido ao maior número de instâncias desta base.

A quantidade de características do melhor indivíduo variou bastante entre o modelo sem aproximação e as variações dos modelos de aproximação. Porém, o nível médio de erro apresentou valores bastante próximos para todos os casos de aproximação, com ligeira vantagem para o modelo por ATRIBUTOS.

Abaixo os gráficos 6.16 a 6.20 com a combinação:

Classificador:.....KNN  
 Aproximação:.....VMP  
 Percentual de Avaliação:.....20%

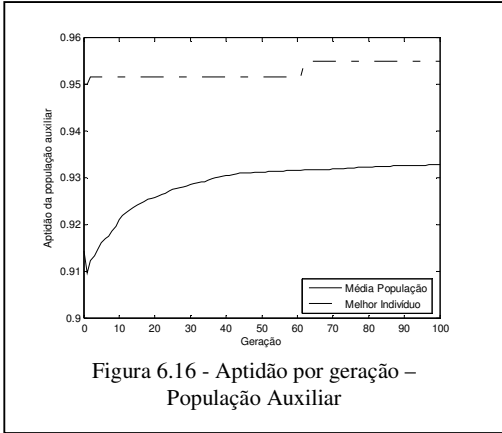


Figura 6.16 - Aptidão por geração – População Auxiliar

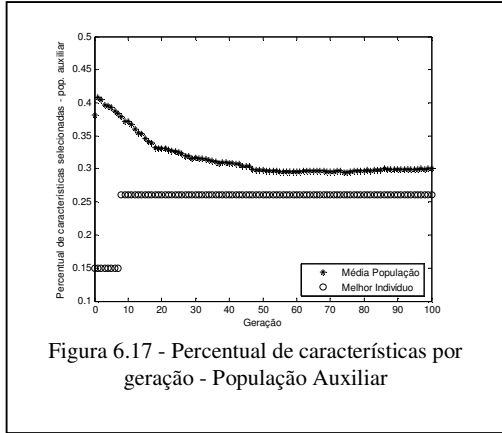


Figura 6.17 - Percentual de características por geração - População Auxiliar

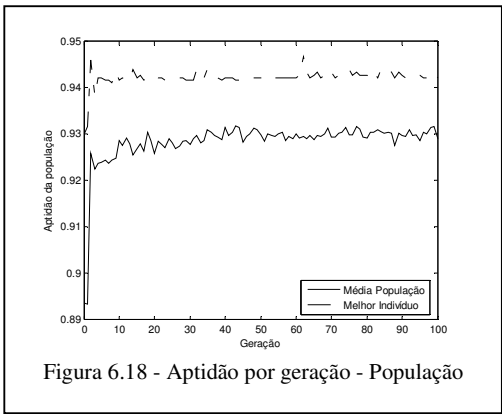


Figura 6.18 - Aptidão por geração - População

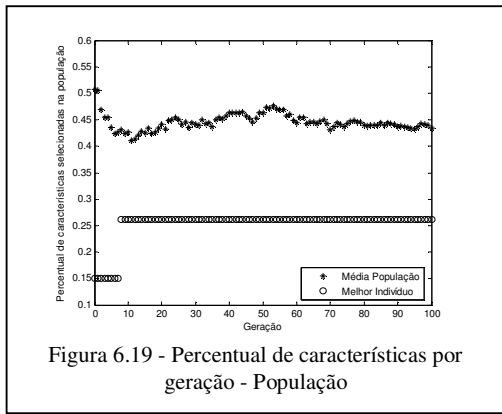


Figura 6.19 - Percentual de características por geração - População

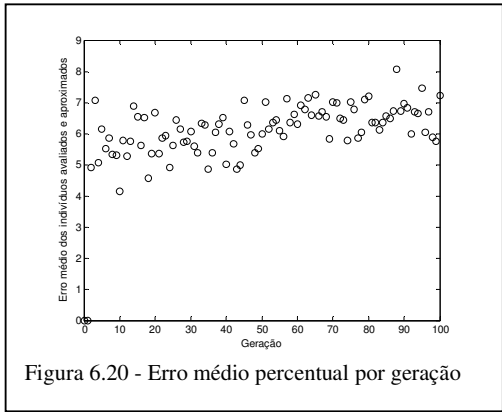


Figura 6.20 - Erro médio percentual por geração

b) Bases Leukemia, Prostate e Colon

Estas bases foram testadas somente com o classificador KNN para o caso sem aproximação e as diversas combinações de aproximação, também em função do tempo de processamento. As tabelas e gráficos adiante mostram os resultados.

Base Leukemia

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Aptidão (%)	Erro médio da aproximação (%)
KNN	SEM	100	223.42	40.0	100.0	-----
	VMP	60	201.91	38.0	100.0	5.79
		40	161.02	43.0	100.0	6.73
		20	119.98	40.0	100.0	7.44
	ATRIBUTOS	60	156.36	37.0	100.0	5.54
		40	119.08	40.0	100.0	5.70
		20	71.81	21.0	100.0	6.79

**Tabela 6.6** Base de Dados Leukemia 72 instâncias e 7130 atributos

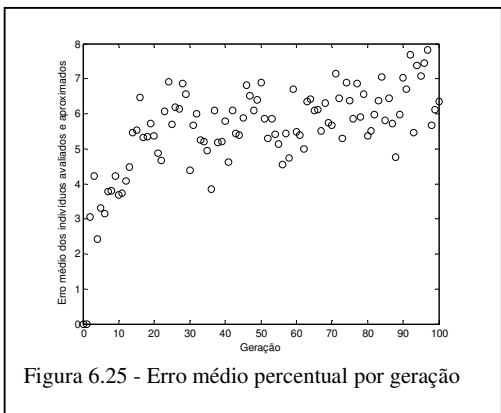
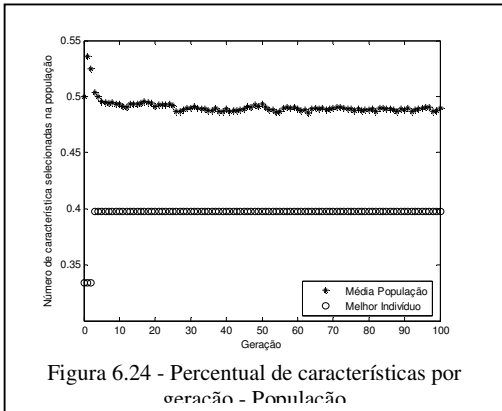
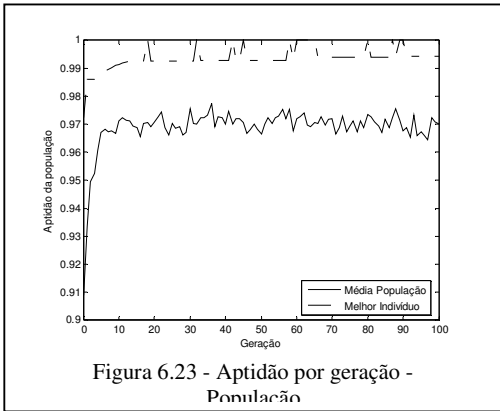
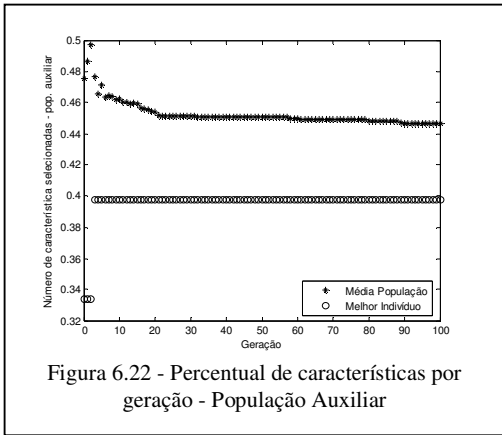
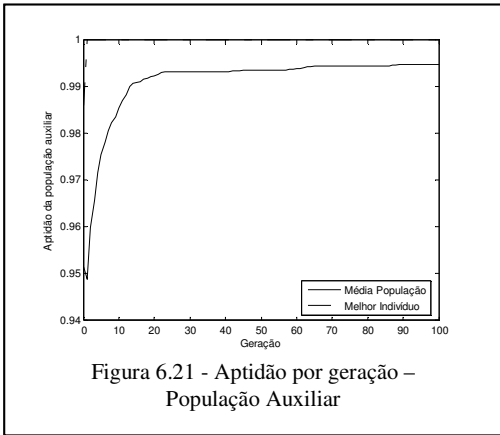
Com esta base, o modelo sem aproximação e todas as alternativas de aproximação resultaram em ótima qualidade de resposta, com 100% de acerto. Porém, o custo computacional, o erro de aproximação e o número de características do melhor indivíduo foram melhores com a técnica de aproximação por ATRIBUTOS.

Abaixo os gráficos 6.21 a 6.25 com a combinação:

Classificador:.....KNN

Aproximação:.....ATRIBUTOS

Percentual de Avaliação:.....40%



## Base Prostate

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Aptidão (%)	Erro médio da aproximação (%)
KNN	SEM	100	749.50	49.0	90.0	-----
	VMP	60	683.53	51.0	94.0	16.20
		40	531.10	45.0	94.0	19.56
		20	389.19	49.0	94.0	14.97
	ATRIBUTOS	60	495.47	46.0	94.0	12.24
		40	375.24	50.0	94.0	12.34
		20	212.13	41.0	91.0	11.46

**Tabela 6.7** - Base de Dados Prostate 102 instâncias e 12600 atributos

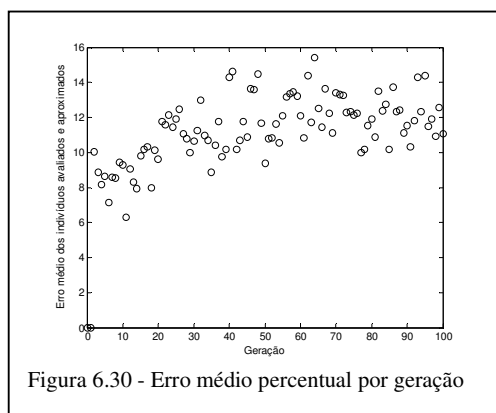
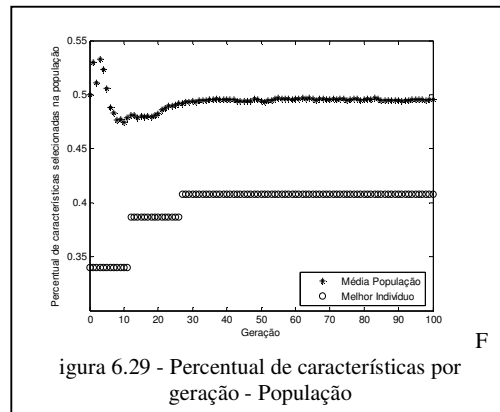
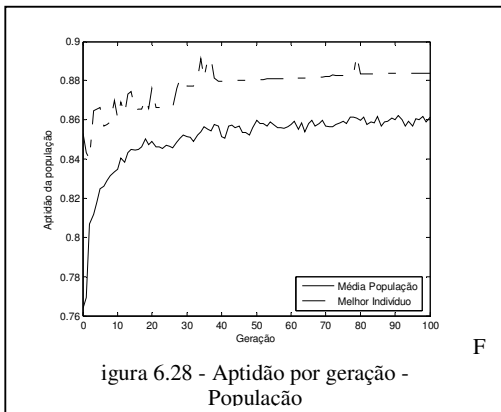
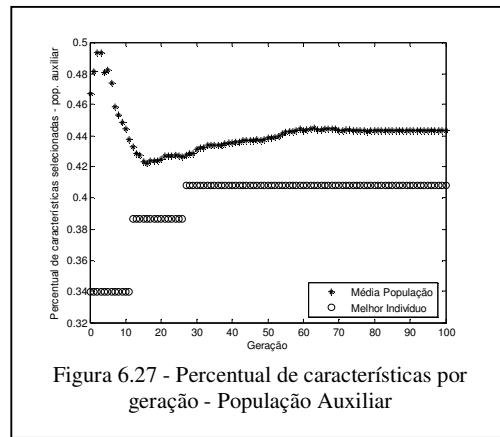
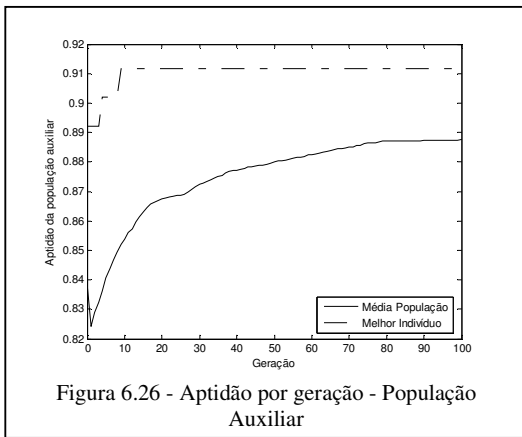
A tabela 6.7 mostra soluções com aptidões equivalentes, porém a aproximação pela técnica de ATRIBUTOS apresenta sempre custos computacionais menores e também erros de aproximação inferiores em relação ao modelo de aproximação VMP.

Abaixo os gráficos 6.26 a 6.30 com a seguinte combinação:

Classificador:.....KNN

Aproximação:.....ATRIBUTOS

Percentual de Avaliação:....20%



## Base Colon

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Aptidão (%)	Erro médio da aproximação (%)
KNN	SEM	100	35.82	43.0	94.0	-----
	VMP	60	47.54	41.0	94.0	7.79
		40	41.41	46.0	94.0	8.27
		20	30.84	40.0	97.0	8.63
	ATRIBUTOS	60	33.74	41.0	97.0	7.17
		40	25.34	41.0	94.0	7.23
		20	15.92	41.0	94.0	6.28

**Tabela 6.8** - Base de Dados Colon 62 instâncias e 2000 atributos

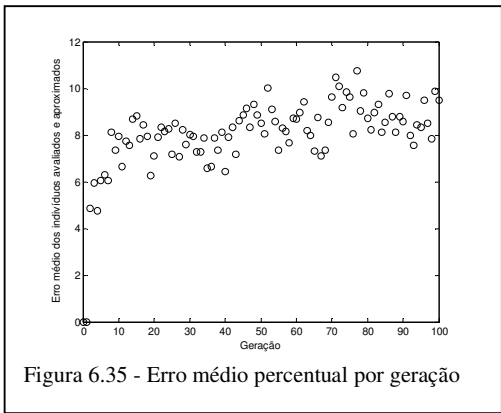
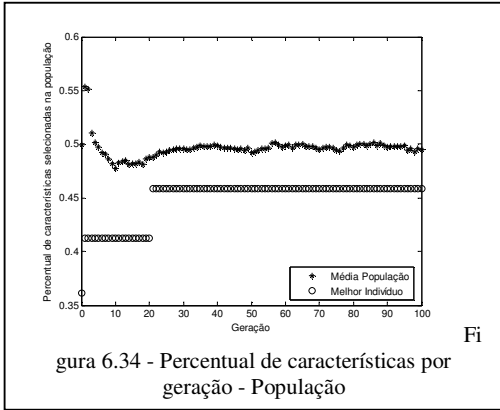
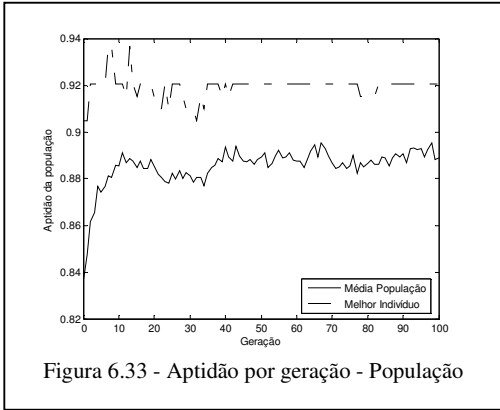
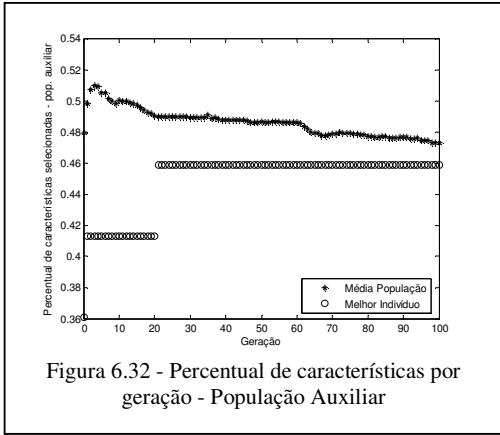
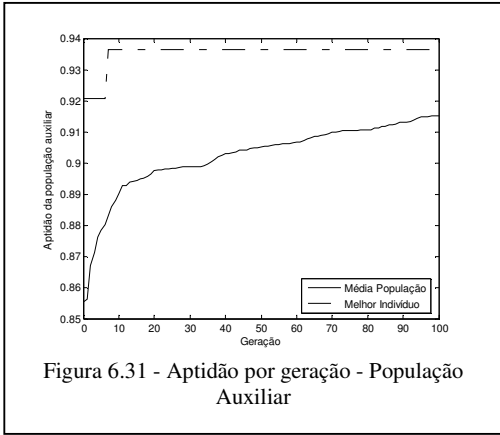
A tabela 6.8, evidencia novamente soluções bem próximas, mas como em outras bases já analisadas, a técnica de aproximação por ATRIBUTOS tem melhor desempenho nos tempos computacionais e no erro médio de aproximação.

Os gráficos 6.31 a 6.35 abaixo são apresentados para a seguinte combinação:

Classificador:.....KNN

Aproximação:.....VMP

Percentual de Avaliação:.....40%



### Base Mushroom

Nesta base de dados não foram feitos os testes com os classificadores KNN e SVM por terem custo computacional muito elevado.

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos dos melhor Indivíduo (%)	Aptidão (%)	Erro médio da aproximação (%)
K-means	SEM	100	719.45	48.0	90.0	-----
	VMP	60	421.07	47.0	91.0	15.00
		40	285.36	46.0	91.0	12.56
		20	145.55	39.0	91.0	14.12
	ATRIBUTOS	60	408.71	41.0	91.0	12.32
		40	256.08	32.0	91.0	11.98
		20	141.85	40.0	91.0	11.28

**Tabela 6.9** Base de Dados Mushroom 5644 instâncias e 98 atributos

Pelos valores encontrados nos testes e mostrados na tabela 6.9 acima, verificamos que as soluções com ambos os métodos de aproximação apresentaram resultados bastante satisfatórios com ligeira vantagem ainda para a técnica de aproximação por ATRIBUTOS em relação ao modelo de aproximação VMP.

Abaixo os gráficos 6.36 a 6.40 com a seguinte combinação:

Classificador:.....K-means

Aproximação:.....ATRIBUTOS

Percentual de Avaliação:.....20%

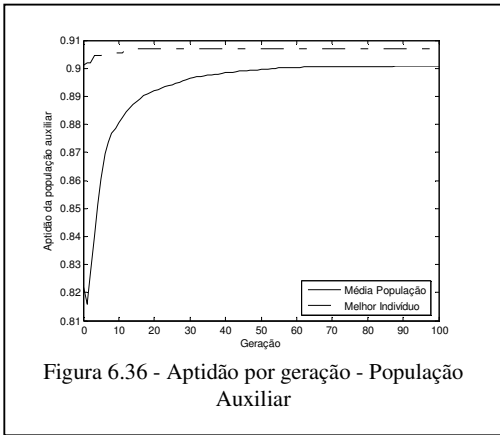


Figura 6.36 - Aptidão por geração - População Auxiliar

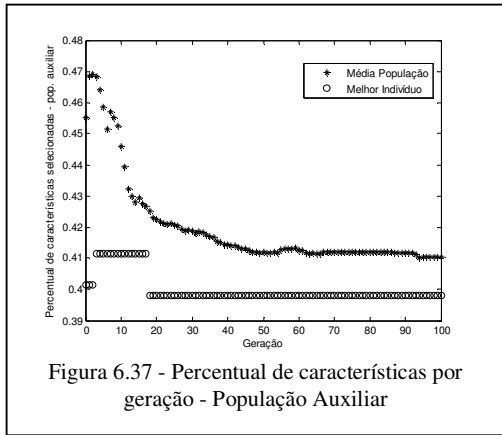


Figura 6.37 - Percentual de características por geração - População Auxiliar

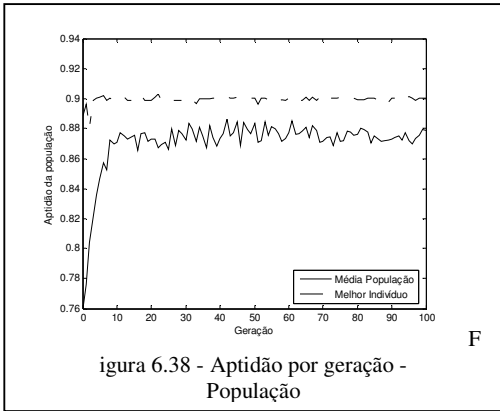


Figura 6.38 - Aptidão por geração - População

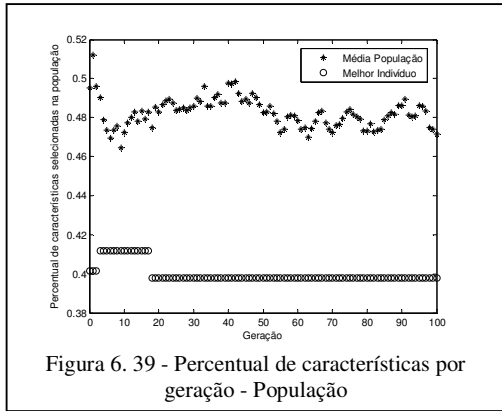


Figura 6.39 - Percentual de características por geração - População

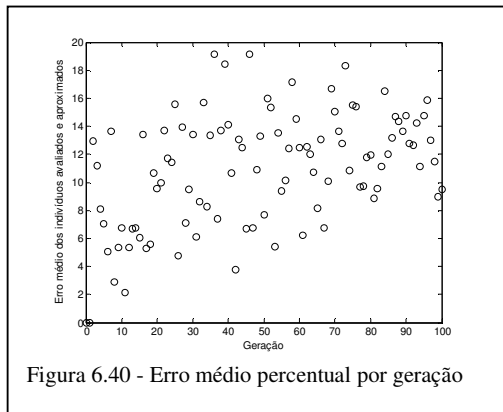


Figura 6.40 - Erro médio percentual por geração

### 6.3.5 Análise dos Gráficos

De uma maneira em geral, o que se percebe nos gráficos é um comportamento uniforme para as bases testadas mostrando o seguinte:

#### **Gráficos de Aptidão**

Uma tendência de melhoria crescente ao longo das gerações, porém com a identificação das melhores soluções nas primeiras gerações.

#### **Gráficos de Características**

Tendência também de, ao longo das gerações, a média de características aproximar-se do número de características da melhor solução.

#### **Gráfico de Erro**

Erro cresce ao longo das gerações, tendendo a se estabilizar em percentuais ainda razoáveis. Porém, como visto nos resultados apresentados, não houve influência na qualidade dos resultados alcançados.

Os gráficos da melhor solução, tanto em aptidão quanto em características, apresentam menor nível de variação. Isto ocorre, principalmente devido à função de aptidão adotada, onde não se tem um componente que priorize soluções que apresentem menor número de características.

## 6.4. Experimentos Adicionais

Duas novas estratégias incorporadas ao modelo de aproximação, a saber, gerenciamento aleatório da seleção de indivíduos a serem avaliados e uma estratégia de penalização a ser utilizada na função objetivo (aptidão) serão consideradas. A descrição e os resultados obtidos, para cada uma das variações, são apresentados a seguir.

- **Gerenciamento aleatório**

Nesta opção substitui-se a seleção determinística para a escolha dos indivíduos a serem avaliados por um procedimento randômico, avaliando o impacto da diminuição da pressão de seleção sobre a qualidade dos resultados em relação ao modelo determinístico. As tabelas e gráficos abaixo mostram estas comparações.

Banco de Dados Sonar

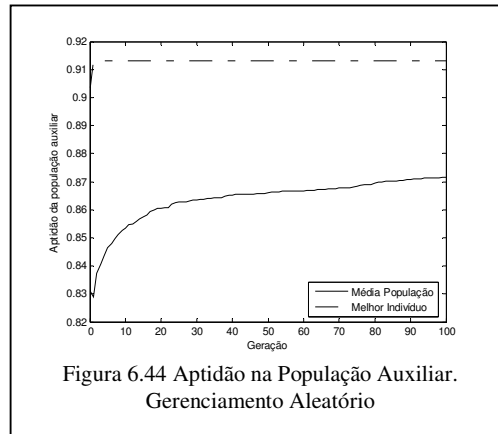
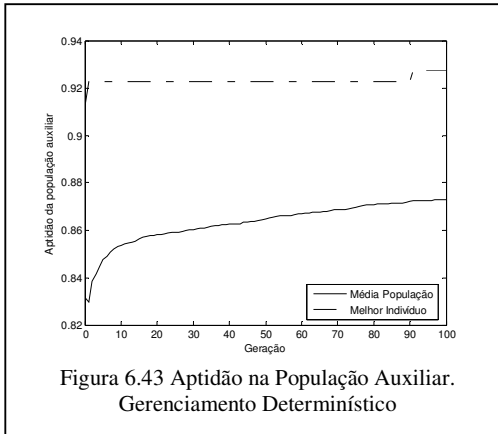
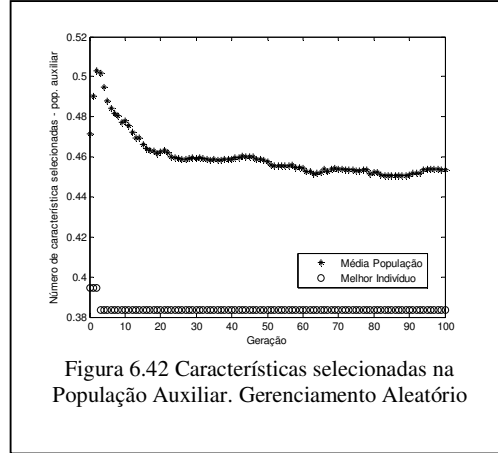
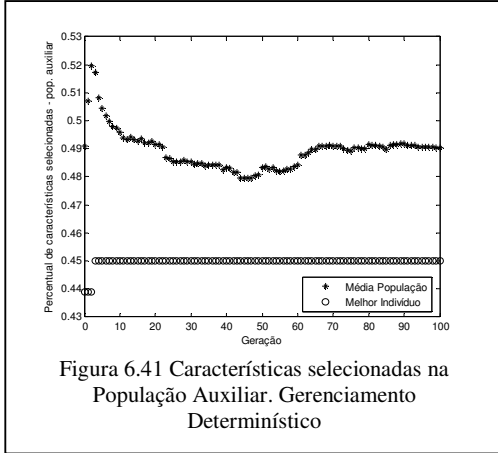
Classificador:.....KNN

Aproximação:.....VMP

Percentual Avaliado:.....20%

Gerenciamento	Tempo Médio Por Execução (seg.)	Características Seleccionadas Melhor Indivíduo	Aptidão Melhor Indivíduo	Erro de Aproximação
Aleatório	5.09	38%	91%	11.30%
Determinístico	4.45	45%	93%	10.70%

**Tabela 6.10** - Base Sonar. Comparação entre gerenciamentos Aleatório e Determinístico



Banco de Dados Ionosphere

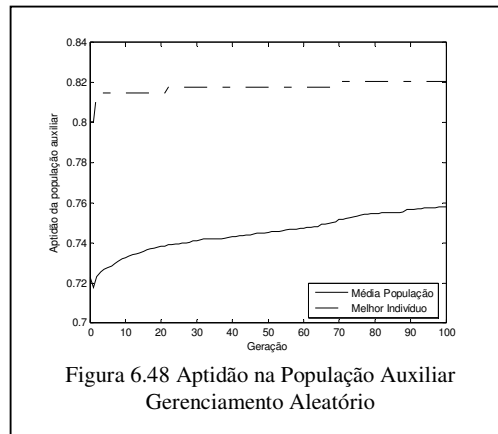
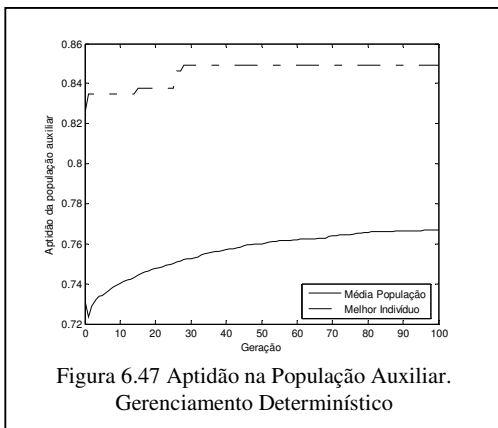
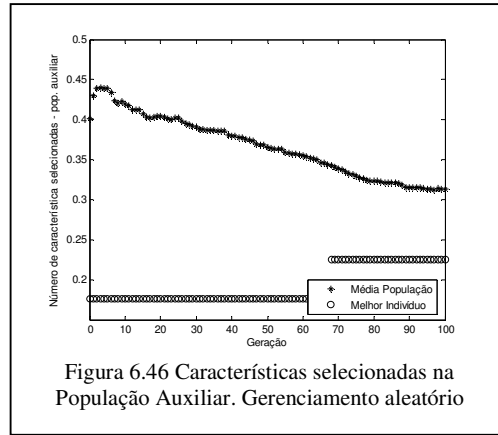
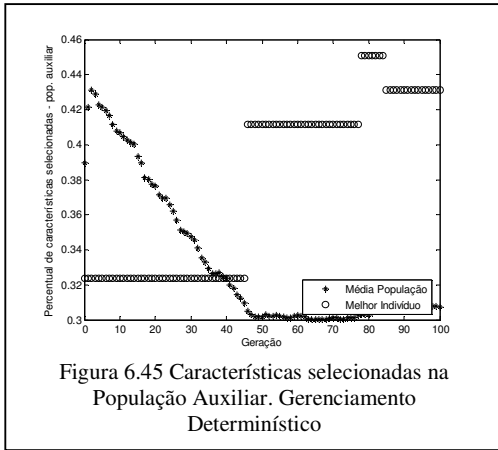
Classificador:..... K-means

Aproximação:.....ATRIBUTOS

Percentual Avaliado:.....20%

Gerenciamento	Tempo Médio Por Execução (seg.)	Características Seleccionadas Melhor Indivíduo	Aptidão Melhor Indivíduo	Erro de Aproximação
Aleatório	6.09	23%	82%	8.97%
Determinístico	6.40	43%	85%	10.70%

Tabela 6.11 - Base Ionosphere. Comparação entre gerenciamentos Aleatório e Determinístico



Os resultados, conforme esperado, apresentaram melhor aptidão com o gerenciamento determinístico, porém a quantidade de características foi menor no gerenciamento aleatório devido à função objetivo (aptidão) adotada.

- **Função com penalização**

Analisa-se a influência de se considerar a função de penalização não-linear, apresentada na seção 5.1.1, dada por:

$$f(x)_{pnl} = (1 + C_{trn(x)}(tst(x))).(2 + nc(x))$$

onde, quanto maior o número de características do indivíduo, maior será a penalidade adotada. Espera-se obter soluções com níveis de acertos similares ao caso sem penalização, porém com um menor número de atributos ativos. A seguir, os resultados:

a) Base de dados Sonar

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Acertos (%)	Erro médio da aproximação (%)	
KNN	SEM	100	25.81	18.0	84.0	-----	
	VMP	60	16.91	18.0	87.0	16.41	
		40	11.56	18.0	88.0	18.01	
		20	6.27	18.0	90.0	19.49	
	ATRIBUTOS	60	17.26	18.0	87.0	12.82	
		40	11.04	18.0	87.0	11.81	
		20	5.93	18.0	93.0	10.00	
	K-means	SEM	100	19.00	27.0	67.0	-----
		VMP	60	15.82	23.0	77.0	22.59
40			8.20	33.0	68.0	18.99	
20			3.99	22.0	74.0	20.39	
ATRIBUTOS		60	12.72	30.0	75.0	16.68	
		40	10.92	20.0	78.0	13.73	
		20	4.82	17.0	74.0	10.10	
SVM		SEM	100	4673.24	18.0	85.0	-----
		VMP	60	2873.49	18.0	93.0	17.14
	40		1998.13	35.0	94.0	21.75	
	20		1019.85	26.0	94.0	21.40	
	ATRIBUTOS	60	2851.71	18.0	94.0	10.58	
		40	1941.80	18.0	94.0	10.86	
		20	1013.96	18.0	94.0	7.86	

**Tabela 6.12** Base Sonar, resultados com penalização

Na tabela 6.12 acima, observa-se que as soluções, para todos os casos analisados, apresentaram um número bem menor de características em relação aos resultados da função sem penalização, apresentada na tabela 6.2. O nível de acerto, em geral, sofreu um pequeno declínio com a utilização da penalização, com o classificador SVM apresentando os melhores resultados.

Os erros médios na aproximação conservaram-se na mesma faixa do modelo que utiliza a função objetivo sem penalização, assim como os esforços computacionais necessários.

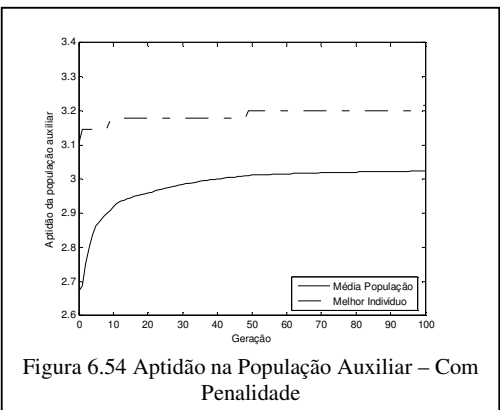
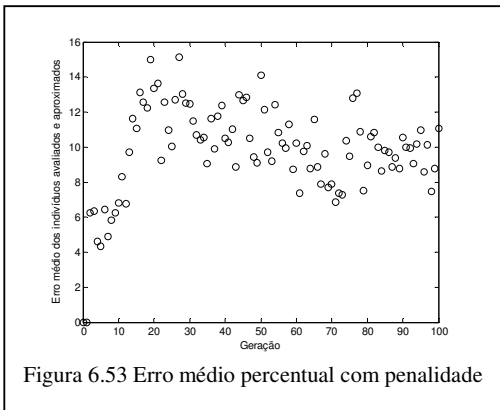
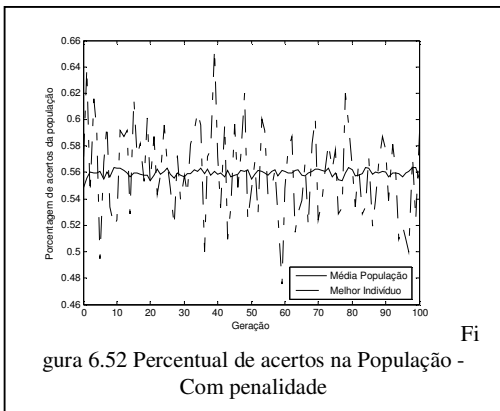
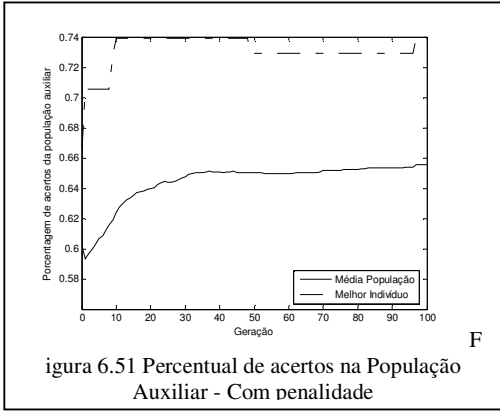
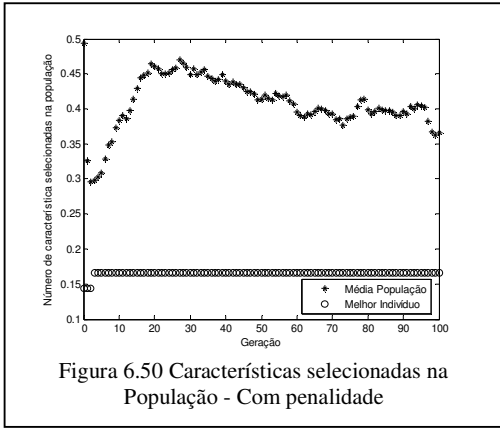
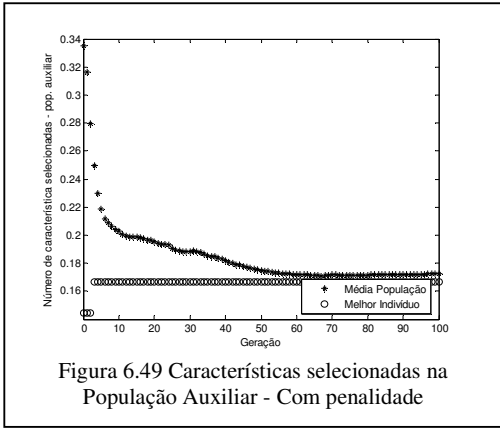
Também, nestes experimentos com penalização, a técnica de aproximação por atributos mostrou, em geral, desempenho melhor. Pode-se considerar que, a função objetivo não-linear adotada apresentou resultados satisfatórios, diminuindo o número de atributos das soluções sem comprometer a qualidade da resposta obtida, tanto para os modelos sem aproximação quanto para os modelos de aproximação testados.

Abaixo os gráficos 6.49 a 6.54, referentes à combinação:

Classificador:.....K-means

Aproximação:.....ATRIBUTOS

Percentual de avaliação:.....20%



Os gráficos para a base Sonar com a função objetivo de penalização não-linear mostram que a evolução do AG em relação ao nível de acerto da população é bastante diferenciada em relação à função sem penalização. O melhor indivíduo varia durante todo o processo com as mudanças constantes no número de características dos indivíduos considerados. Porém, o gráfico da função objetivo (aptidão) mostra um comportamento mais uniforme, com evolução mais suave e contínua tanto do melhor indivíduo quanto da média da população auxiliar. Deve-se observar que, por construção, a função objetivo com penalização sempre apresenta valor maior que a unidade.

b) Base de dados Breast

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Acertos (%)	Erro médio da aproximação (%)	
KNN	SEM	100	63.84	20.0	88.0	-----	
	VMP	60	236.87	20.0	100.0	27.10	
		40	213.26	20.0	100.0	28.48	
		20	195.20	20.0	100.0	29.39	
	ATRIBUTOS	60	79.79	20.0	100.0	16.50	
		40	66.10	20.0	100.0	16.39	
		20	50.66	20.0	100.0	14.46	
	K-means	SEM	100	167.77	20.0	83.0	-----
		VMP	60	282.90	20.0	100.0	33.69
40			279.52	19.0	100.0	38.90	
20			242.91	34.0	100.0	28.09	
ATRIBUTOS		60	157.53	20.0	100.0	26.90	
		40	120.42	20.0	100.0	26.83	
		20	81.36	24.0	100.0	24.06	
SVM		SEM	100	412.45	20.0	83.0	-----
		VMP	60	417.89	20.0	100.0	33.86
	40		349.94	20.0	100.0	36.05	
	20		264.78	20.0	100.0	37.02	
	ATRIBUTOS	60	272.81	20.0	100.0	16.80	
		40	195.83	20.0	100.0	16.62	
		20	113.08	20.0	100.0	14.07	

**Tabela 6.13** Base Breast, resultados com penalização

Nesta base, a função objetivo com penalidade apresentou pouca influência nos resultados em relação aos resultados obtidos com a função sem penalização, mostrados na tabela 6.3. Apesar da quantidade de atributos e das poucas instâncias, esta base não mostrou maiores dificuldades para a obtenção de classificadores com resultados ótimos, tanto para o modelo sem aproximação tanto para as variações dos modelos com aproximação utilizados, o que tornou, de certa forma, a penalização menos efetiva.

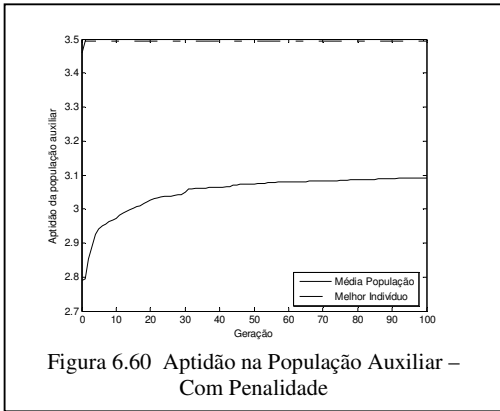
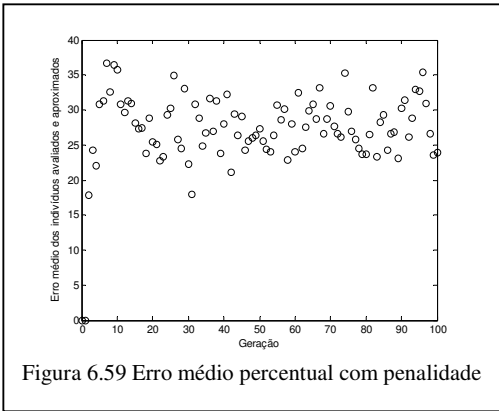
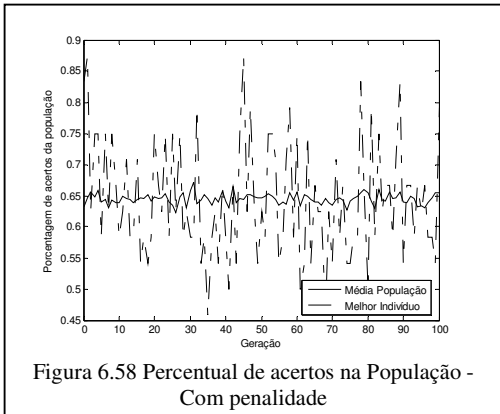
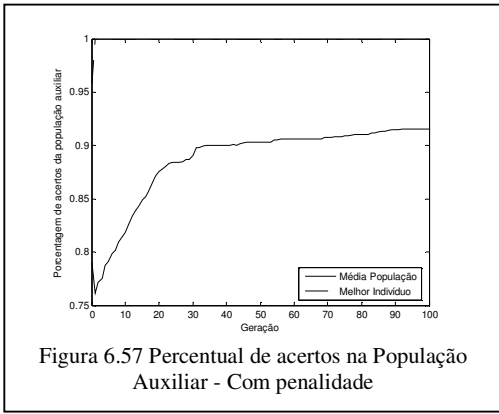
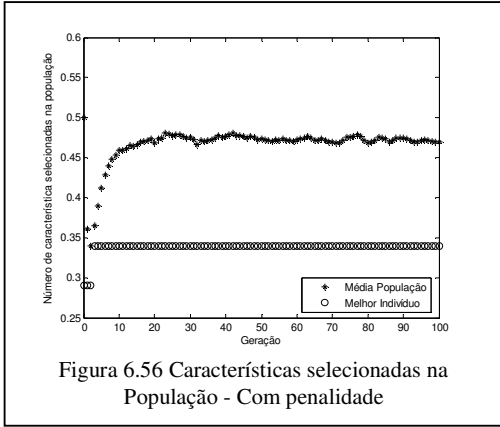
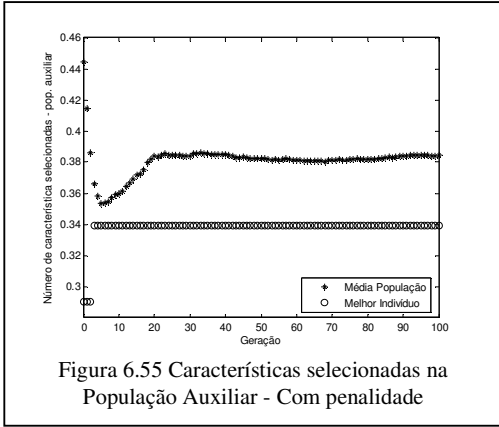
Ressalta-se o menor nível de erro de aproximação e menor custo computacional do modelo de aproximação por ATRIBUTOS em relação ao modelo VMP.

Abaixo, os gráficos 6.55 a 6.60, referentes à combinação:

Classificador:.....K-means

Aproximação:.....VMP.

Percentual de avaliação:....20%.



c) Base de dados Ionosphere

Classificador	Aproximação	Percentual Avaliado	Tempo médio por execução	Atributos do melhor Indivíduo (%)	Acertos (%)	Erro médio da aproximação (%)
KNN	SEM	100	38.66	6.00	91.0	-----
	VMP	60	24.03	6.00	93.0	17.26
		40	17.98	8.00	89.0	19.93
		20	9.56	8.00	91.0	22.03
	ATRIBUTOS	60	26.03	6.00	93.0	13.78
		40	16.73	6.00	94.0	12.89
		20	8.49	8.00	91.0	6.04
K-means	SEM	100	22.95	10.00	76.0	-----
	VMP	60	8.89	11.00	80.0	19.56
		40	7.56	20.00	82.0	21.64
		20	5.03	11.00	80.0	18.47
	ATRIBUTOS	60	12.69	11.00	82.0	11.50
		40	9.71	12.00	81.0	13.30
		20	5.29	9.00	81.0	8.47
SVM	SEM	100	15751.93	12.00	91.0	-----
	VMP	60				
		40				
		20	3353.01	15.0	92.0	14.97
	ATRIBUTOS	60				
		40				
		20	3374.13	10.0	91.0	4.27

**Tabela 6.14** Base Ionosphere, resultados com penalização

A tabela 6.14 indica que a aplicação da função objetivo com penalização neste banco de dados apresentou um excelente resultado em relação à função sem penalização, mostrado na tabela 6.4. Verifica-se que o número de características das soluções obtidas caiu drasticamente, sem perda no nível de acerto alcançado.

Tanto na qualidade da resposta quanto em relação ao tempo de execução, os resultados com o modelo VMP e a o modelo por ATRIBUTOS foram bem similares. Porém, a aproximação por ATRIBUTOS apresentou um erro médio na aproximação bem inferior ao modelo VMP.

Abaixo, os gráficos 6.61 a 6.66, referentes à combinação:

Classificador:.....K-means

Aproximação:.....VMP.

Percentual de avaliação:...40%.

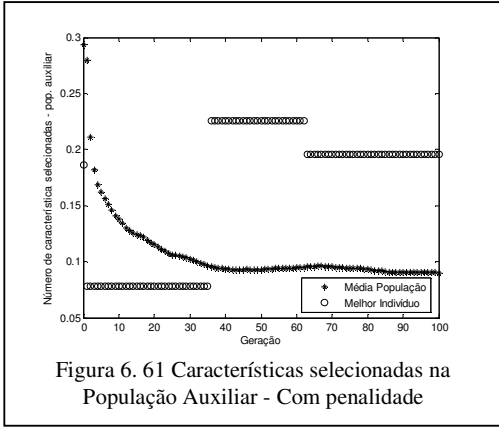


Figura 6.61 Características selecionadas na População Auxiliar - Com penalidade

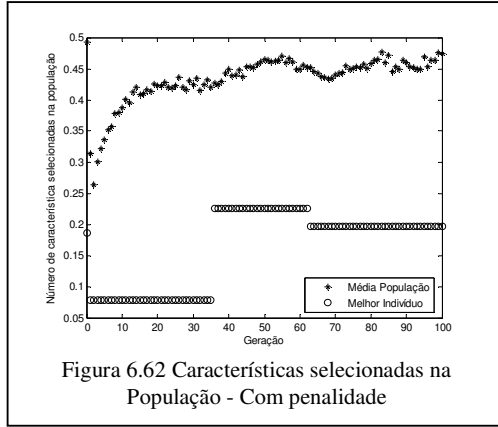


Figura 6.62 Características selecionadas na População - Com penalidade

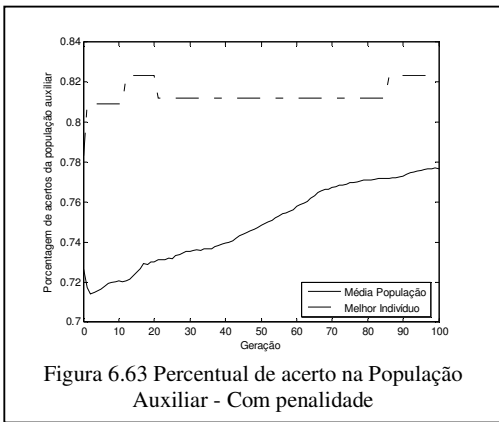


Figura 6.63 Percentual de acerto na População Auxiliar - Com penalidade

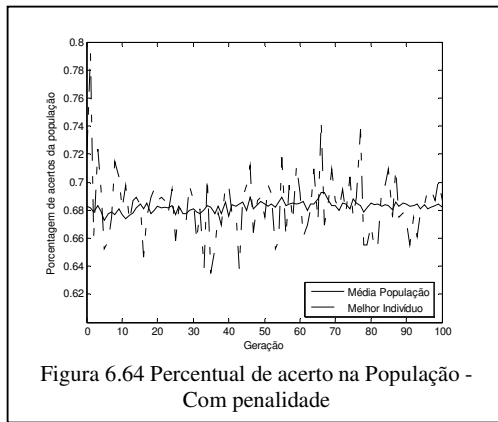


Figura 6.64 Percentual de acerto na População - Com penalidade

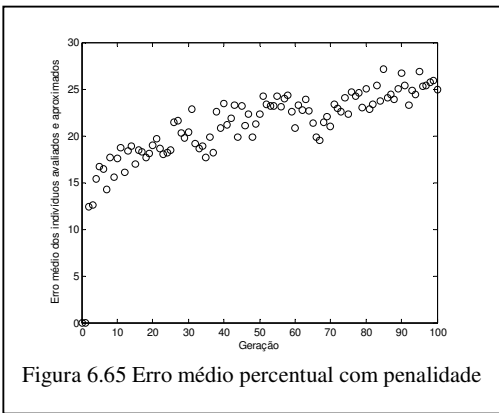


Figura 6.65 Erro médio percentual com penalidade

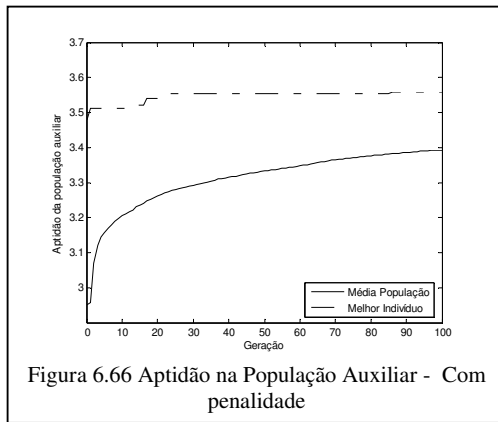


Figura 6.66 Aptidão na População Auxiliar - Com penalidade

## CAPÍTULO 7 Conclusões e Perspectivas

Apresentam-se, a seguir, as principais conclusões obtidas com a aplicação dos modelos propostos neste trabalho, assim como os possíveis e previstos direcionamentos visando dar continuidade a esta linha de pesquisa.

Estratégias de seleção de características encapsuladas tem se mostrado mais eficientes em relação aos modelos de seleção baseado em filtro. Porém, o esforço computacional dos procedimentos encapsulados pode tornar proibitiva sua aplicação.

Um modelo de aproximação construído aproveitando-se das propriedades específicas de algoritmos evolucionistas, a saber, algoritmos genéticos, é proposto visando viabilizar a seleção de características encapsulada com qualidade similar ao modelo sem aproximação e com custo computacional reduzido.

Entre as principais características do modelo de aproximação proposto estão:

- Heurística específica para geração da população inicial;
- Gerenciamento com escolha determinística dos indivíduos a serem avaliados e atualização efetiva da população auxiliar;
- Definição de funções objetivo sem a penalização e penalizando soluções com excesso de atributos;
- Cálculo da avaliação aproximada através de um modelo proposto, baseado nos alelos (atributos) da candidata a solução em codificação binária;

Experimentos numéricos realizados com classificadores diferenciados e bases de dados heterogêneas apresentaram resultados bastante satisfatórios. O gerenciamento determinístico mostrou eficiência considerável em relação a um modelo randômico, indicando que o custo adicional para a aproximação de indivíduos que deveriam ser avaliados é recompensado por resultados mais atraentes.

O formato adotado para a atualização por geração e determinística da base de dados de referência para a aproximação, também chamada de população auxiliar, mostrou-se interessante por ser efetivo no cálculo da aproximação e por ser um repositório de indivíduos mais robustos do que a população do algoritmo evolucionista.

A função objetivo não-linear de penalização indicou ser viável no sentido de forçar a obtenção de soluções que apresentem um número reduzido de características com perdas mínimas na acurácia dos classificadores obtidos.

Em geral, os resultados com os modelos de aproximação testados obtiveram resultados de certa forma surpreendentes, com níveis de predição superiores ao alcançados quando não se utiliza a aproximação. Acredita-se que isto aconteça devido ao processo de seleção dos indivíduos a serem avaliados. Tanto a aproximação por vizinho mais próximo quanto à aproximação por alelos determinam os indivíduos a serem avaliados (melhores) através de uma seleção baseada nas variáveis de projeto (atributos) e não somente através do espaço das funções objetivo, como é feito normalmente pelos algoritmos evolucionistas. Tal processo indica ser eficaz no direcionamento da busca do subconjunto ótimo de características.

Em relação aos classificadores utilizados (SVM, KNN e K-means) concluiu-se que foi mantido o padrão de acurácia e custo computacional de cada um deles com a utilização do modelo de aproximação. Neste sentido, nenhuma anomalia foi detectada.

Porém, pode-se considerar que o modelo de aproximação por alelo ou atributo apresentou o desempenho mais favorável da técnica como um todo. Desenvolvido para aplicação em codificações com baixa epistasia adequou-se perfeitamente ao perfil geralmente utilizado para codificar características para o processo de seleção de atributos, obtendo soluções com qualidade e custo computacional compatíveis com a tradicional técnica de vizinhos mais próximos.

Pretende-se, em continuidade ao trabalho, estudar com maior profundidade o efeito da epistasia sobre o modelo de aproximação por atributo, aplicando-o em funções com alta epistasia para verificação de seu desempenho. Deve-se também, investigar mais detalhadamente a influência da seleção dos indivíduos a serem avaliados pelo modelo aproximado na qualidade da resposta obtida.

Estudos do comportamento das técnicas de aproximação em problemas de classificação com múltiplas classes e em bancos desbalanceados estão previstos. No caso das bases que apresentam desbalanceamento entre as classes, possivelmente novas funções objetivo serão necessárias porque, neste caso, a porcentagem de acertos deixa de ser uma medida confiável da qualidade do classificador.

A aplicação do modelo de aproximação em outros problemas como, por exemplo, problemas de otimização estrutural também é de grande interesse, pois sua codificação pode apresentar níveis variados de epistasia.

Finalmente, ressalta-se a possibilidade de desenvolvimento de um modelo de aproximação híbrido, para computação evolucionista, onde o gerenciador seria responsável por coordenar as simulações, as aproximações por atributo e, também, aproximações por herança.

## REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL S.C., **Metamodeling: a study of approximations in queueing models**. MITPress, Cambridge, MA, USA, 1985.

AKBARZADEH-T M.R., DAVARYNEJAD M., PARIZ N., Adaptive fuzzy fitness granulation for evolutionary optimization. *International Journal of Approximate Reasoning*, v.49, n.3, pp.523–538, 2008.

ALTMAN N.S., An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* v.46, n.3, pp.175–185, 1992.

ARIS R., **Mathematical modelling techniques**. Courier Dover Publications, USA, 1994.

BÄCK T., Selective pressure in evolutionary algorithms: a characterization of selection mechanisms. *Proceedings of The First IEEE Conference on Evolutionary Computation*, IEEE Computer Society Press, v.1, pp.57–62, 1994.

BARBOSA H., A coevolutionary genetic algorithm for constrained optimization problems. *Congress on Evolutionary Computation*. Washington, USA, v.3, pp.1605–1611, 1999.

BARTON R.R., Simulation metamodels. *WSC'98: Proceedings of the 30<sup>th</sup> conference on Winter Simulation*, IEEE Computer Society Press, pp.167–176, 1998.

BLANNING R.W., The source and uses of sensitivity information. *Interfaces* 4(4), pp. 21–23, 1974.

BLANNING R.W., Response to Michel, Kleijnen and Permut. *Interfaces* 5(3), pp.24–25, 1975.

BLUM A.L. e LANGLEY P., Selection of relevant features and examples in machine learning. *Artificial Intelligence*, pp.245-271, 1997.

BORGES C.C.H., **Algoritmos genéticos para otimização em dinâmica de estruturas**. Tese de Doutorado, PEC-COPPE/UFRJ, Rio de Janeiro, Brasil, 1999.

BUECHE D., SCHRAUDOLPH N., KOUMOUTSAKOS P., Accelerating evolutionary algorithms with gaussian process fitness function models. *IEEE Trans on Systems, Man, and Cybernetics: Part C* 35(2), pp.183–194.2005.

CARVALHO D.R., **Árvore de decisão/algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados**. Tese de Doutorado, PEC-COPPE/UFRJ, Rio de Janeiro, Brasil, 2005.

CHEN V.C.P., TSUI K.L., BARTON R.R. et al, A review on design, modeling and applications of computer experiments. IEEE Transactions v.38 n.4 pp.273–291, 2006.

DARWIN C. R., **On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.** John Murray, 2<sup>a</sup>. Ed., London, 1860. Disponível em <http://darwin-online.org.uk/contents.html>. Último acesso em 05/06/2009.

DASH M. e LIU H., Feature selection for classification. Intelligence Data Analysis - An International Journal v.1, n.3, 1997.

DOAK J., **An evaluation of features selection methods and their application to computer security.** Technical report, Davis CA. University of California, Department of Computer Science, 1992.

DUCHEYNE E., DE BAETS B., DE WULF R., **Is fitness inheritance useful for real-world applications?** Second International Conference on Multi-criterion Optimization, Springer, LNCS 2632, pp.31–42, 2003.

EMMERICH M., GIANNAKOGLU K., NAUJOKS B., Single and multiobjective evolutionary optimization assisted by gaussian random field metamodels. Evolutionary Computation v.10, n.4, pp.421–439, 2006.

FERRARI S., STENGEL R.F., Smooth function approximation using neural networks. IEEE Transactions on Neural Networks v.16, n.1, pp.24–38, 2005.

FOGEL D.B., An introduction to simulated evolutionary optimization. IEEE Transactions on Neural Networks v.5, n.1, pp.3–14, 1994.

FONSECA L.G., **Algoritmos genéticos assistidos por metamodelos baseados em similaridade.** Tese de Doutorado, LNCC, Petrópolis, Brasil, 2009.

FONSECA L.G., BARBOSA H.J.C., LEMONGE A.C.C., Metamodel assisted genetic algorithm for truss weight minimization. ICMOSPS'07, Durban, South Africa, CD-ROM, 2007.

FORRES, S., and MITCHELL, M. Relative building-block fitness and the building-block hypothesis. In D. Whitley (ed.), Foundations of Genetic Algorithms 2 Morgan Kaufman, SanMateo, CA, 1993.

GOLDBERG D., **Genetic algorithms in search, optimization and machine learning.** Addison-Wesley Publishing Co., reading, Mass., USA, 1989.

GORISSEN D., Towards an adaptive and flexible metamodeling toolbox. Relatório Técnico TR-06-14, COMS Research Group, University of Antwerp, 2006.

GORISSEN D., HENDRICKX W., CROMBECQ K., DHAENE T., Integrating gridcomputing and metamodeling. CCGRID '06: Proceedings of the Sixth IEEE

International Symposium on Cluster Computing and theGrid(CCGRID'06), IEEE Computer Society, Washington, DC, USA, pp.185–192, 2006.

GREFENSTETTE J., FITZPATRICK J., Genetic search with approximate fitness evaluations. Proceedings of the International Conference on Genetic Algorithms and Their Applications, pp.112–120, 1985.

GUYON I. Welcome and introduction to the problem of feature/variable selection. NIPS 2001 Workshop on Feature/variable selection, 2001.

HASTIE T., TIBSHIRANI R. e FRIEDMAN J., **The elements of statistical learning**. Springer, 2001.

HENDRICKX W., DHAENE T., Sequential design and rational metamodelling. WSC '05: Proceedings of the 37th conference on Winter simulation, Winter Simulation Conference, pp.290–298, 2005.

HOLLAND J.H., Adaptation in natural and artificial system. Ann Arbor, The University of Michigan Press, 1975.

HUSSAIN M.F., BARTON R.R., JOSHI S.B., Metamodeling: radial basis functions, versus polynomials. European Journal of Operational Research, 2002.

JACOBS J.H., ETMAN L.F.P., VAN KEULEN F., ROCHA J.E., Framework for sequential approximate optimization. Structural and multidisciplinary optimization 27, pp.384–400, 2004.

JAIN A. e ZONGKER D., Feature selection: evaluation,application and small sample performance. IEEE Trans. Pattern Analysis and Machine Intelligence, v.19, n.2, pp. 153-158, 1997.

JIN Y., A comprehensive survey of fitness approximation in evolutionary computation. Soft Computing Journal v.9, n.1, pp.3–12, 2005.

JOHN G.H., **Enhancements to the data mining process**. PhD Thesis, Department of Computer Science, Stanford University, 1997.

JOHN G.H.,KOHAVI R., PFLEGER K., Irrelevant features and the subset selection problem. In Proceedings of the Eleventh International Conference on Machine Learning, pp.121-129, New Brunswick, NJ, 1994.

KAUFFMAN, S.A. **The origins of order, self-organization and selection in evolution**. Oxford University Press, New-York , 1993.

KECMAN V., **Learning and soft computing: support vector machines, neural networks, and fuzzy logic models**. MIT Press, Cambridge, MA, USA, 2001.

KIM H.S., CHO S.B., An efficient genetic algorithm with less fitness evaluation by clustering. *Evolutionary Computation*, 2001 Proceedings of the 2001 Congress v.2, pp. 887–894, 2001.

KLEIJNEN J.P.C., SARGENT R.G., A methodology of fitting and validating metamodels in simulation. *European Journal of Operational Research* v.120, pp.14–29, 2000.

KRÖSE B.J.A., VAN DER SMAGT P.P., **An introduction to neural networks**. University of Amsterdam, Amsterdam, 1993.

KYBIC J, BLU T, UNSER M., Generalized sampling: a variational approach - Part I: theory. *IEEE Transactions on Signal Processing*, v.50, n.8, pp.1965–1976, 2002.

LEE W., STOLFO S.J. AND MOK K.W., Adaptive intrusion detection: a data mining approach. *AI Review*, v.14, n.6, pp.533-567, 2000.

LESH F.H., Multi-dimensional least-squares polynomial curve fitting. *Commun ACM* 2(9), pp.29–30, 1959.

LEOPOLD L. AND KINDERMANN J., Text categorization with support vector machines. How to represent texts in input space? *Machine Learning*, 46, pp.423-444, 2002.

LIU H. AND MOTODA H., **Feature selection for knowledge discovery and data mining**. Boston, Kluwer Academic Publishers, 1998.

LIU H. AND SETIONO R., A probabilistic approach to feature selection: a filter solution. In *Proc. Of the 13<sup>th</sup> Int. Conf. of Machine Learning*, pp.319-327, Morgan Kaufmann, 1996.

MENDONÇA, C. E. L. R., **Um sistema computacional para otimização através de algoritmos genéticos e redes neurais**. Tese de Doutorado, PEC-COPPE, UFRJ, Rio de Janeiro, Brasil, 2004.

MITCHELL T.M., **Machine learning**. McGraw-Hill, New York, 1997.

MOCCIARDI M., A comparison of seven techniques of choosing subsets of pattern recognition. *IEEE Trans. Computers*, C.20, pp.1023-1031, 1971.

MOTA F., GOMIDE F., Fuzzy clustering in fitness estimation models for genetic algorithms and applications. *2006 IEEE International Conference on Fuzzy Systems*, pp.1388–1395, ISBN: 0-7803-9488-7, 2006.

MULLUR A.A., MESSAC. A., Metamodeling using extended radial basis functions: a comparative approach. *Engineering with Computers* 21(3), pp.203–217, 2006.

NARENDRA P.M. AND FUKUNAGA K., A branch and bound algorithm for feature subset selection. *IEEE Trans. On Computer*, C-26(9), pp.917-922, 1977.

NIGAM K., MCCALLUM A.K., THRUN S. AND MITCHELL T., Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 46, pp.103-134, 2000.

PEREIRA, S. C. A., **Tratamento de incertezas em modelagens de bacias**. Tese de Doutorado, PEC-COPPE, UFRJ, Rio de Janeiro, Brasil, 2002.

QUACKENBUSH J., Computational analysis of microarray data. The Institute for Genomic Research, 9, 712 Medical Center Drive, Rockville, Maryland 20850 USA.

RUNARSSON T., Approximate evolution strategy using stochastic ranking. Em: Yen GG, Wang L, Bonissone P, Lucas SM (eds) *IEEE World Congress on Computational Intelligence*, Vancouver, Canada, 2006.

RUI Y., HUANG T.S. AND CHANG S., Image retrieval: current techniques, promising directions and open issues. *Visual Communication and Image Representation*, 10(4), pp.39-62, 1999.

SARAIVA, J. M. F., **A utilização de redes neurais em conjunto com o método de Monte Carlo na análise da confiabilidade de estruturas**. Tese de Doutorado, PEC-COPPE, UFRJ, Rio de Janeiro, Brasil, 1997.

SASTRY K., **Evaluation-relaxation schemes for genetic and evolutionary algorithms**. Dissertação de Mestrado, University of Illinois at Urbana-Champaign, IL, USA, Urbana, IL, 2002.

SAUNDERS G., GAMMERMAN A., VOVK V., Ridge regression learning algorithm in dual variables. *Proc. 15th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp.515–521, 1998.

SHEPARD D. A., Two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 23rd ACM National Conference*, ACM Press, New York, NY, USA, pp.517–524, 1968.

SIEDLECKI W. AND SKANSKY., On automatic feature selection. *International journal of Pattern Recognition and Artificial Intelligence*, 2, pp.197-220, 1988.

SIMPSON T.W., POPLINSKI J., KOCH P.N., ALLEN J., Metamodels for computer-based engineering design: survey and recommendations. *Engineering with Computers* 17(2), pp.129–150, 2001.

SIRONEN S., KANGAS A., MALTAMO M., KALLIOVIRTA J., Localization of growth estimates using non-parametric imputation methods. *Forest Ecology and Management* v.256, pp.674–684, 2008.

SMITH R.E, DIKE B.A., STEGMANN S.A., Fitness inheritance in genetic algorithms. *SAC '95: Proceedings of the 1995 ACM symposium on Applied computing*, ACM Press, New York, NY, USA, pp.345–350, 1995.

SWETS D.L. AND WENG J.J., Efficient content-based image retrieval using automatic feature selection. In IEEE International Symposium on Computer Vision, pp. 85-90, 1995.

TAYLOR M.K., AUCLAIR P.F., MYKYTKA E.F., Working smarter when developing linear simulation metamodels. WSC '95: Proceedings of the 27<sup>th</sup> conference on Winter simulation, pp.1392–1399, 1995.

ULMER H., STREICHER F., ZELL A., Model-assisted steady-state evolution strategies. Proceedings of Genetic and Evolutionary Computation Conference, LNCS 2723, pp.610–621, 2003a.

ULMER H., STREICHER F., ZELL A., Evolution strategies assisted by gaussian processes with improved pre-selection criterion. Proceedings of IEEE Congress on Evolutionary Computation, pp.692–699, 2003b.

VAN BEERS W.C.M, KLEIJNEN J.P.C., Kriging interpolation in simulation: a survey. WSC '04: Proceedings of the 36th Conference on Winter simulation, Winter Simulation Conference, pp.113–121, 2004.

VAPNIK, V.N., CHERVONENKIS, A., On the uniform convergence of relative frequencies of events to their probabilities. Theoretical Probability and Its Applications, vol.17, pp.264-280, 1971.

VAPNIK, V. M., **The nature of statistical learning theory**. Springer-Verlag, New York Inc, New York, 1995.

VAPNIK, V. M., **Statistical learning theory**. John Wiley and Sons, New York, 1998.

YANG Y. AND PEDERSON J.O., A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, pp.412-420, 1997.