UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EVERTON FERNANDO BARROS

Um algoritmo cultural para descoberta de conhecimento em banco de dados

Maringá
2011

EVERTON FERNANDO BARROS

Um algoritmo cultural para descoberta de conhecimento em banco de dados

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ademir Aparecido Constantino

Maringá
2011

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

B277a

Barros, Everton Fernando
Um Algoritmo cultural para descoberta de conhecimento em banco de dados / Everton Fernando Barros. -- Maringá, 2011.
169 f. : il. (algumas col.), figs., tabs.

Orientador: Prof. Dr. Ademir Aparecido Constantino.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Informática, Programa de Pós-Graduação em Ciência da Computação, 2011.

1. Mineração de dados. 2. Programação genética. 3. Algoritmo cultural - Programação genética. Plano de saúde - Mineração de dados. 4. Knowledge Discovery in Databases (KDD). 5. Mineração de dados - Impressões gerais. 6. Algoritmo evolucionário. I. Constantino, Ademir Aparecido, orient. II. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22.ed. 006.312

AMMA-00047

FOLHA DE APROVAÇÃO

EVERTON FERNANDO BARROS

Um algoritmo cultural para descoberta de conhecimento em banco de dados

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

Prof. Dr. Ademir Aparecido Constantino
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Wesley Romão
Universidade Estadual de Maringá – DIN/UEM

Profa. Dra. Valéria Delisandra Feltrim
Universidade Estadual de Maringá – DIN/UEM

Profa. Dra. Gisele Lobo Pappa
Universidade Federal de Minas Gerais – DCC/UFMG

Aprovada em: 02 de dezembro de 2011.
Local da defesa: Sala 101, Bloco C56, *campus* da Universidade Estadual de Maringá

DEDICATÓRIA(S)

Dedico esta dissertação aos meus pais e aos meus irmãos por sempre me incentivar e apoiar a estudar e a lutar pelos meus sonhos.

AGRADECIMENTO(S)

Ao meu orientador e professor Dr. Ademir Aparecido Constantino pela atenção e dedicação em sanar minhas dúvidas e pela sua orientação.

Ao professor Dr. Wesley Romão pelos muitos conselhos, dicas e orientações ao longo do trabalho que me proporcionaram chegar a conclusão do mesmo.

Aos meus pais Mario Aparecido Barros e Clair Balan Barros, pelo exemplo de luta, pelo incentivo em todos os momentos de minha vida e por sempre estarem dispostos a ajudar.

Aos meus irmãos Jean Fábio Barros e Gustavo Henrique Barros, pelo apoio.

A todos os professores do programa que de alguma maneira contribuíram para a realização desse trabalho.

Á Maria Inês Davanço, secretária do mestrado que sempre foi muito prestativa.

A todos os amigos do mestrado em especial Gustavo Sato, Marcelo R. Borth (Borth), Alexandre Huff, Mauro H. Mulati, José Magon Junior, Murilo Zangari de Souza, Lurdete Fontaniva, Rafael Vivian, Juliana Keiko Yamaguchi, Filipe Roseiro Côgo, Vinícius M. A. de Souza e César Alberto da Silva, pelos momentos de apoio e descontração.

Aos amigos que não são do mestrado, André Luiz Sales de Moraes, Elison José Gracite Lusvardi, Wilson Guerra, Flávio José da Silva.

Aos funcionários da Benner Maringá pela cooperação, apoio e ajuda ao longo da dissertação em especial ao Celso Lara de Souza que cooperou de maneira direta com a pesquisa.

Ao apoio financeiro da CAPES e ao Programa de Pós-Graduação em Ciência da Computação (PCC-UEM).

Um algoritmo cultural para descoberta de conhecimento em banco de dados

RESUMO

Os planos de saúde, tanto públicos quanto privados, vem acumulando grande quantidade de dados contendo informações ocultas que poderiam ajudar seus gestores na redução de custos das operadoras de planos de saúde e auxiliar no planejamento de programas de prevenção de doenças. Esses conhecimentos podem ser descobertos utilizando técnicas de Mineração de Dados (MD) para extração de informações relevantes e úteis. Nesse contexto pretende-se resolver a tarefa de classificação, que permite a descoberta de regras de previsão. Nesta dissertação apresenta-se uma abordagem para resolver a tarefa de classificação em MD na qual se propõe um algoritmo híbrido que utiliza Programação Genética (PG) juntamente com Algoritmo Cultural (AC), que é um algoritmo evolucionário baseado no processo de evolução cultural da humanidade. A abordagem implementada teve como estudo de caso os dados de uma operadora de planos de saúde suplementar, contendo informações administrativas e de procedimentos em hospitais, laboratórios e consultórios, relativos a beneficiários do estado de Santa Catarina. Esses dados foram pré-processados e preparados para serem utilizados pelo algoritmo proposto. Para avaliar a abordagem apresentada foram realizados experimentos com PG e com AC, onde o AC armazena conhecimentos e ajuda a guiar o processo evolutivo. Um desses conhecimentos armazena as impressões gerais, que são as crenças do usuário, permitindo medir o interesse nas regras encontradas. Com base nestes experimentos foi avaliada a capacidade do algoritmo proposto em encontrar regras em duas situações: considerando as impressões gerais do usuário e sem considerar as impressões gerais (crenças do usuário) e obteve-se bons resultados pelas duas abordagens. Avaliou-se também a capacidade do algoritmo em encontrar regras a partir de um conjunto de dados contendo pequenos disjuntos no qual conseguiu-se encontrar regras tanto dos grandes disjuntos como dos pequenos disjuntos, além de avaliar a maneira como as impressões gerais do usuário podem ser utilizadas pelo AC.

Palavras-chave: Mineração de dados, Programação Genética, Algoritmos Culturais, Plano de Saúde.

A cultural algorithm for knowledge discovery in databases

ABSTRACT

Health plans, both public and private, have been accumulating large amounts of data containing hidden information that could help their managers to reduce costs of health insurance carriers and assist in planning programs for disease prevention. This knowledge can be discovered using Data Mining (DM) techniques for extracting information relevant and useful. In this context we intend to solve the task of classification that allows the discovery of prediction rules. This dissertation presents an approach to solving the classification task in MD which proposes a hybrid algorithm that uses Genetic Programming (GP) with Cultural Algorithm (CA), which is an evolutionary algorithm based on the process of cultural evolution of humanity. The approach was implemented as a case study data from an operator of health insurance supplement, containing administrative information and procedures in hospitals, laboratories and offices, for the beneficiaries of the state of Santa Catarina. These data were preprocessed and prepared for use by the algorithm. Experiments were performed to evaluate the approach with GP and CA, where CA stores knowledge and helps guide the evolutionary process. A knowledge of these stores general impressions, which are the beliefs of the user, allowing to measure the interest in the rules found. Based on these experiments we evaluated the ability of the algorithm to find rules in two situations: considering the general impressions of the user and without considering the general impression (the user's beliefs) and obtained good results by the two approaches. We also evaluated the ability of the algorithm to find rules from a data set containing small disjoint in which it was possible to find rules which belong to small disjoint and big disjoint, and assess how the general impression of the user can be used for CA.

***Keywords*:** Data Mining, Genetic Program, Cultural Algorithms, Health Insurance.

LISTA DE FIGURAS

Figura 2.1 Uma visão dos passos do processo de KDD. (Fayyad *et al.*, 1996a) . 24
Figura 2.2 Métodos de Criação de intervalos. Fonte: (Han e Kamber, 2006) . . 28
Figura 2.3 Um gráfico 2-D sobre dados de clientes e suas localizações em uma cidade, demonstrando três agrupamentos. Cada centroide é marcado com um "+", representando o ponto médio no espaço do grupo. Valores extremos podem ser detectados como valores fora do conjunto de grupos. Fonte: (Han e Kamber, 2006) 29
Figura 2.4 Estrutura de uma árvore de decisão. Fonte: (Coelho, 2004) 33
Figura 2.5 Uma matriz de confusão para exemplos positivos e negativos 36
Figura 2.6 Comparação da regra com IG. Fonte: (Romão, 2002) 38
Figura 2.7 Separação dos Dados 39
Figura 2.8 Validação Cruzada com 4 partições 40
Figura 2.9 Representação de um indivíduo (programa) em PG 42
Figura 2.10 Algoritmo PTC1 44
Figura 2.11 Dois indivíduos antes do cruzamento. 47
Figura 2.12 Dois filhos produzidos pelo cruzamento. 47
Figura 2.13 Ação do operador de mutação *point mutation* em um indivíduo. . . 48
Figura 2.14 Algoritmo Cultural (Reynolds *et al.*, 2003) 53
Figura 2.15 Funcionamento básico de um algoritmo cultural. 53
Figura 2.16 Representação do Conhecimento Situacional (Reynolds e Saleem, 2001). 56
Figura 2.17 Representação do Conhecimento Normativo (Reynolds e Saleem, 2001). 57
Figura 2.18 Representação do Conhecimento Histórico (Becerra e Coello, 2005). 58

Figura 3.1 Conjuntos difusos para três valores linguísticos 65
Figura 3.2 Representação do Indivíduo 67
Figura 3.3 Indivíduo Semanticamente Inválido 68
Figura 3.4 Seleção de subárvores dos pais 69
Figura 3.5 Troca de subárvores dos pais gerando novos indivíduos 70
Figura 3.6 Mutação sobre um nó interno (função) 70
Figura 3.7 Mutação sobre um nó folha (terminal) 71
Figura 3.8 Matriz de confusão difusa 72
Figura 3.9 Função de pertinência (FP) para Consulta = "Baixo" 73
Figura 3.10 Função de pertinência para SADT = "Alto" 73

Figura 3.11 Exemplo de matriz de confusão difusa 74
Figura 3.12 Fluxograma do algoritmo PG 78

Figura 4.1 Algoritmo Cultural Implementado 81
Figura 4.2 Fluxograma do Algoritmo Cultural Desenvolvido 82
Figura 4.3 Atualização do Conhecimento Situacional 85
Figura 4.4 Influência do Conhecimento Situacional no cruzamento 86
Figura 4.5 Exemplo de mapa para o Conhecimento Topográfico 87
Figura 4.6 Atualização do Conhecimento Topográfico 88
Figura 4.7 Influência do Conhecimento Topográfico no cruzamento 89
Figura 4.8 Influência do Conhecimento Topográfico 90
Figura 4.9 Indivíduos gerados a partir de IG's 91
Figura 4.10 Influência do Conhecimento de Domínio no cruzamento 91

Figura 5.1 Frequência do grupo CID XXI em relação aos demais grupos 107
Figura 5.2 Histograma de frequência de agrupamentos de códigos de CID . . . 107
Figura 5.3 Frequência do grupo CID XXI em relação aos demais grupos desconsiderando a quantidade de atendimentos 108
Figura 5.4 Histograma de frequência de agrupamentos de códigos de CID desconsiderando repetições 109
Figura 5.5 Frequência de Clínica Médica em relação as demais especialidades . 109
Figura 5.6 Frequência de Clínica Médica em relação as demais especialidades desconsiderando repetições 110
Figura 5.7 Frequência de SADT, Consulta e Internação 111
Figura 5.8 Frequência de SADT, Consulta e Internação sem repetições 111
Figura 5.9 Frequência do atributo Sexo 112
Figura 5.10 Frequência de Grupos de Eventos 113
Figura 5.11 Frequência de Grupos de Eventos sem repetições 113

Figura 6.1 Gráfico de Pareto para os efeitos padronizados do GRUPO-CID-IIIA 123
Figura 6.2 Gráfico dos efeitos principais para o GRUPO-CID-IIIA 124
Figura 6.3 Gráfico de Pareto para os efeitos padronizados do GRUPO-CID-IIIB 124
Figura 6.4 Gráfico dos efeitos principais para o GRUPO-CID-IIIB 125
Figura 6.5 Gráfico de Pareto para os efeitos padronizados do GRUPO-EVENTO-09 125
Figura 6.6 Gráfico dos efeitos principais para GRUPO-EVENTO-09 126
Figura 6.7 Evolução da taxa de acerto e interesse para GRUPO-CID-IIIB (AC1) 132

Figura 6.8 Evolução da média da taxa de acerto e interesse para GRUPO--CID-IIIB (AC1) 132
Figura 6.9 Ajuste das taxas de mutação e cruzamento para GRUPO-CID--IIIB (AC1) no decorrer do processo evolutivo 133
Figura 6.10 Ajuste das probabilidades dos conhecimentos para GRUPO-CID-IIIB (AC1) no decorrer do processo evolutivo 133
Figura 6.11 Evolução da aptidão para GRUPO-CID-IIIB (AC1) 133
Figura 6.12 Evolução da média de aptidão da população para GRUPO-CID--IIIB (AC1) 134
Figura 6.13 Evolução da taxa de acerto para GRUPO-CID-IIIB (AC2) 135
Figura 6.14 Evolução da média da taxa de acerto para GRUPO-CID-IIIB (AC2) 135
Figura 6.15 Ajuste das taxas de mutação e cruzamento para GRUPO-CID--IIIB (AC2) no decorrer do processo evolutivo 136
Figura 6.16 Ajuste das probabilidades dos conhecimentos para GRUPO-CID-IIIB (AC2) no decorrer do processo evolutivo 136
Figura 6.17 Evolução da aptidão para GRUPO-CID-IIIB (AC2) 137
Figura 6.18 Evolução da média de aptidão para GRUPO-CID-IIIB (AC2) . . . 137

LISTA DE TABELAS

Tabela 3.1 Conjunto de dados fictício 73

Tabela 5.1 Tabela de Beneficiário, GRS-BENEFICIARIO 104
Tabela 5.2 Tabela de Guia, GRS-GUIA 104
Tabela 5.3 Tabela de Guia Evento, GRS-GUIA-EVENTO 104
Tabela 5.4 Tabela Geral de Eventos, SAM-TGE 104
Tabela 5.5 Tabela de códigos de CID, SAM-CID 104
Tabela 5.6 Descrição dos atributos selecionados 114
Tabela 5.6 Descrição dos atributos selecionados 115
Tabela 5.6 Descrição dos atributos selecionados 116

Tabela 6.1 Atributos selecionados 119
Tabela 6.2 Distribuição de classes 119
Tabela 6.3 Fatores, Níveis e Variações do Projeto de Experimentos 120
Tabela 6.4 Matriz com o planejamento dos experimentos 121
Tabela 6.5 Execução do Experimento para a Classe "GRUPO-CID-IIIA = SIM" 122
Tabela 6.6 Execução do Experimento para a Classe "GRUPO-CID-IIIB = SIM" 122
Tabela 6.7 Execução do Experimento para a Classe "GRUPO-EVENTO-09 = SIM" 123
Tabela 6.8 Valores dos parâmetros para cada classe 126
Tabela 6.9 Taxa de acerto utilizando validação cruzada para k = 6 (PG) . . . 127
Tabela 6.10 Taxa de acerto PG1 x PG2 x C5.0 127
Tabela 6.11 Cobertura PG1 x PG2 x C5.0 127
Tabela 6.12 Média do número de antecedentes PG1 x PG2 x C5.0 128
Tabela 6.13 Impressões Gerais (GRUPO-CID-IIIA) 128
Tabela 6.13 Impressões Gerais (GRUPO-CID-IIIA) 129
Tabela 6.14 Impressões Gerais (GRUPO-CID-IIIB) 129
Tabela 6.15 Impressões Gerais (GRUPO-EVENTO-09) 130
Tabela 6.16 Taxa de acerto utilizando validação cruzada para k = 6 (AC1) . . . 131
Tabela 6.17 Taxa de acerto utilizando validação cruzada para k = 6 (AC2) . . . 134
Tabela 6.18 Taxa de acerto AC1 x AC2 137
Tabela 6.19 Interesse AC1 x AC2 137
Tabela 6.20 Cobertura AC1 x AC2 138

Tabela 6.21 Média do número de antecedentes AC1 x AC2 138
Tabela 6.22 Regras para o atributo GRUPO-CID-IIIA (PG) 138
Tabela 6.22 Regras para o atributo GRUPO-CID-IIIA (PG) 139
Tabela 6.23 Regras para o atributo GRUPO-CID-IIIB (PG) 139
Tabela 6.23 Regras para o atributo GRUPO-CID-IIIB (PG) 140
Tabela 6.24 Regras para o atributo GRUPO-EVENTO-09 (PG) 140
Tabela 6.24 Regras para o atributo GRUPO-EVENTO-09 (PG) 141
Tabela 6.25 Regras para o atributo GRUPO-CID-IIIA (AC1) 142
Tabela 6.26 Regras para o atributo GRUPO-CID-IIIB (AC1) 142
Tabela 6.26 Regras para o atributo GRUPO-CID-IIIB (AC1) 143
Tabela 6.27 Regras para o atributo GRUPO-EVENTO-09 (AC1) 143
Tabela 6.27 Regras para o atributo GRUPO-EVENTO-09 (AC1) 144
Tabela 6.28 Regras para o atributo GRUPO-CID-IIIA (AC2) 144
Tabela 6.28 Regras para o atributo GRUPO-CID-IIIA (AC2) 145
Tabela 6.29 Regras para o atributo GRUPO-CID-IIIB (AC2) 145
Tabela 6.29 Regras para o atributo GRUPO-CID-IIIB (AC2) 146
Tabela 6.30 Regras para o atributo GRUPO-EVENTO-09 (AC2) 146
Tabela 6.30 Regras para o atributo GRUPO-EVENTO-09 (AC2) 147

LISTA DE ABREVIATURAS E SIGLAS

AC	Algoritmo Cultural
AE	Algoritmo Evolucionário
AG	Algoritmo Genético
ANS	Agência Nacional de Saúde Suplementar
CID	Classificação Internacional de Doenças
DCBD	Descoberta de Conhecimento em Banco de Dados
FND	Forma Normal Disjuntiva
FP	Função de Pertinência
GPEA	Grupo de Pesquisa em Engenharia de Algoritmo
IG	Impressões Gerais
KDD	*Knowledge Discovery in Databases*
MD	Mineração de Dados
MS	Ministério da Saúde
PG	Programação Genética
PSS	Plano de Saúde Suplementar
PTC	*Probabilist Tree-Creation*
SADT	Serviço Auxiliar Diagnóstico e Terapia
TRS	Terapia Renal Substitutiva
UEM	Universidade Estadual de Maringá

SUMÁRIO

1 Introdução **17**

1.1 Considerações Iniciais 17
1.2 Contexto e Motivação 18
1.3 Objetivos 19
1.4 Organização do Trabalho 20

2 Revisão Bibliográfica **21**

2.1 Considerações Iniciais 21
2.2 Descoberta de Conhecimento em Banco de Dados (DCBD) 21
2.2.1 Etapas do processo de DCBD 22
2.2.2 Pré-Processamento 24
2.2.3 Mineração de Dados 32
2.3 Programação Genética 40
2.3.1 Representação dos programas 41
2.3.2 Fechamento e Suficiência 42
2.3.3 População Inicial 42
2.3.4 Função de Aptidão 45
2.3.5 Métodos de Seleção 45
2.3.6 Operadores Genéticos 46
2.3.7 Critério de Término 49
2.3.8 Um algoritmo básico de PG 49
2.4 Algoritmos Culturais 51
2.4.1 Inspiração Natural 51
2.4.2 Funcionamento Básico de um AC 52
2.4.3 Principais Característica 54
2.4.4 Microevolução x Macroevolução 55
2.4.5 Tipos de Conhecimento 56
2.5 Mineração de dados em Planos de Saúde 58
2.5.1 Avaliação da eficácia do tratamento 59
2.5.2 Gestão em Saúde 59
2.5.3 Gestão de Relacionamento com o Beneficiário 60
2.5.4 Detecção de Fraudes e Abuso 61
2.6 Considerações Finais 62

3 Programação Genética Aplicada à Mineração de Dados 63
3.1 Considerações Iniciais . . . 63
3.2 Modelo do Sistema . . . 63
3.2.1 Funções e Terminais . . . 64
3.2.2 Representação do Indivíduo . . . 66
3.2.3 Operadores Genéticos . . . 69
3.2.4 Elitismo . . . 71
3.2.5 Avaliação das Regras (Função *Fitness*) . . . 71
3.2.6 Fluxograma do Sistema . . . 77
3.3 Considerações Finais . . . 78

4 Algoritmo Cultural Aplicado à Mineração de Dados 80
4.1 Considerações Iniciais . . . 80
4.2 Modelo do Sistema . . . 81
4.2.1 Inicialização, Avaliação da População e Criação do Espaço de Crenças 81
4.2.2 Geração e Seleção de Indivíduos e Atualização do Espaço de Crenças 82
4.3 Espaço Populacional . . . 83
4.4 Espaço de crenças . . . 83
4.5 Tipos de Conhecimento . . . 84
4.5.1 Conhecimento Situacional . . . 84
4.5.2 Conhecimento Topográfico . . . 86
4.5.3 Conhecimento De Domínio . . . 89
4.5.4 O problema da mutação nos conhecimentos . . . 91
4.6 Interesse nas Regras . . . 92
4.6.1 Cálculo da Similaridade do Antecedente da Regra . . . 93
4.6.2 Cálculo do Grau de Interesse . . . 94
4.6.3 Cálculo da aptidão do Indivíduo . . . 95
4.7 Adaptação dos Parâmetros . . . 96
4.8 Considerações Finais . . . 99

5 Preparação de Dados 100
5.1 Considerações Iniciais . . . 100
5.2 Métodos . . . 101
5.2.1 Entendimento do Domínio da Aplicação . . . 102
5.2.2 Seleção de um Conjunto de Dados Alvo . . . 102
5.2.3 Limpeza dos dados . . . 103

5.2.4 Pré-seleção de atributos . . . 103
5.2.5 Redução e projeção de dados . . . 106
5.3 Considerações Finais . . . 116

6 Experimentos e Resultados **118**
6.1 Considerações Iniciais . . . 118
6.2 Descrição Geral dos Experimentos . . . 118
6.3 Primeiro Experimento . . . 119
6.3.1 Definição do Problema . . . 120
6.3.2 Escolha de Fatores, Níveis e Variações . . . 120
6.3.3 Seleção da Variável de Resposta . . . 120
6.3.4 Escolha do Projeto Experimental . . . 120
6.3.5 Execução do Experimento . . . 121
6.3.6 Análise Estatística dos Dados . . . 121
6.3.7 Conclusões e Recomendações . . . 126
6.3.8 Resultados (PG) . . . 126
6.3.9 Comparação de Resultados . . . 127
6.4 Segundo Experimento . . . 128
6.4.1 Experimento com interesse durante o processo evolutivo (AC1) . . . 130
6.4.2 Experimento com interesse após o processo evolutivo (AC2) . . . 134
6.4.3 Comparação de Resultados . . . 136
6.5 Regras Encontradas . . . 138
6.5.1 Regras PG . . . 138
6.5.2 Regras AC1 . . . 142
6.5.3 Regras AC2 . . . 144
6.6 Considerações Finais . . . 147

7 Conclusões e Trabalhos Futuros **149**
7.1 Conclusões . . . 149
7.2 Trabalhos Futuros . . . 150

Referências **152**

Apêndice A Descrição dos atributos dos dados de saúde **160**

Apêndice B Tabela de frequência dos atributos de especialidade **165**

Apêndice C Distribuição de classes dos dados de saúde **168**

Capítulo

1

Introdução

1.1 Considerações Iniciais

O armazenamento de dados em sistemas computacionais já se tornou uma prática comum e essencial, principalmente devido à queda do custo de armazenamento dos dados e a rápida automatização das empresas, que faz com que a quantidade de dados armazenados em bases de dados cresça a uma velocidade alta. Essa grande quantidade de dados existe em muitos domínios, tais como: financeiro, comercial, científico, produção, manufatura e médico (Rezende, 2003).

Essa grande quantidade de dados pode também esconder conhecimento precioso. Muito desse conhecimento está oculto e é potencialmente útil, entretanto não pode ser descoberto por meio de análises manuais. Além disso, a quantidade de dados disponíveis nessas bases está muito além da compreensibilidade humana necessitando-se o uso de computadores. A análise de grandes volumes de dados e a transformação de dados em conhecimento serve, em última análise, para apoiar a tomada de decisões em diversos segmentos.

Para transformar esses dados em conhecimentos Fayyad *et al.* (1996b) unificou o processo para descoberta de conhecimento em banco de dados e definiu o processo de KDD (*knowledge discovery in databases*) em nove etapas, nas quais *mineração de dados* refere-se a um passo particular nesse processo. Seguindo a definição proposta por Fayyad no que diz respeito a etapa de mineração de dados, e visando melhorar o apoio aos gestores, essa etapa tem sido aplicada aos dados para explorar suas relações de causa e efeito, no

qual o maior desafio é torná-los úteis, extraindo deles informações corretas, compreensíveis e interessantes.

Na área de assistência à saúde, é crescente a quantidade de dados que os gestores de planos de saúde tanto público quanto privados vem acumulando. Esses dados podem conter pepitas de conhecimento importantes para redução de custos e programas de prevenção de doenças. Para buscar essas pepitas em geral é necessário aplicar técnicas inteligentes que agilizem o processo de extração de conhecimentos relevantes. Portanto, as técnicas de descoberta de conhecimento e mineração de dados podem apoiar gestores de planos de saúde na tomada de decisões.

1.2 Contexto e Motivação

As técnicas de mineração de dados (MD) podem ser aplicadas para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis e relevantes que poderiam permanecer ocultos e ignorados. A etapa de MD geralmente é dividida em duas categorias principais: as tarefas de previsão e as tarefas descritivas. Nesta dissertação pretende-se abordar uma das tarefas de previsão, que é a tarefa de classificação, a qual consiste na previsão de um valor categórico baseado nos valores predecessores das instâncias. Por exemplo, prever se um indivíduo tem predisposição para uma doença específica. Para prever valores categóricos utilizou-se programação genética (PG) juntamente com algoritmos culturais (AC).

Os algoritmos culturais são baseados na ideia de que a cultura evolui ao longo das gerações e foi proposto por Reynolds (1994). Além disso, a cultura evolui mais rapidamente que a evolução genética e isso permite uma melhor adaptação ao ambiente do que a possível adaptação pela genética. Essa melhor adaptação ao ambiente e evolução mais veloz indicam o potencial dos algoritmos culturais para resolver a tarefa de classificação em MD de forma mais eficiente que a abordagem evolucionária simples. É importante salientar que os algoritmos culturais precisam de um algoritmo que faça uso de população de indivíduos para funcionar, e portanto, podem ser modelados sobre qualquer técnica de inteligência artificial que utilize tal abordagem. Para preencher essa necessidade do AC optou-se pela PG pois trabalha com populações de indivíduos, além de outras características interessantes como a utilização de funções na representação dos indivíduos que foram importantes para escolha do algoritmo.

A PG pertence ao ramo da computação evolucionária que se baseia no princípio da seleção natural (Banzhaf *et al.*, 1998; Koza, 1992). Alguns autores consideram a PG como uma variação dos algoritmos genéticos, entretanto, há algumas diferenças entre essas duas

técnicas. Uma das diferenças que despertou o interesse por essa técnica, em detrimento de outros algoritmos que fazem uso de população de indivíduos, é a representação do indivíduo, que na maioria dos casos encontrados na literatura é representado utilizando estrutura de árvores. Essa representação mostrou-se interessante quando utilizadas para resolver a tarefa de classificação (Cavaretta e Chellapilla, 1999).

É importante que o conhecimento descoberto seja interessante para o usuário. Para isso existem técnicas que utilizam conhecimento prévio do usuário. Uma dessas técnicas utiliza impressões gerais, que são um tipo de conhecimento específico que o usuário tem a respeito do domínio da aplicação. Essa técnica pode ser utilizada juntamente com PG e AC, pois podem ajudar na descoberta de conhecimento relevante.

Um protótipo foi implementado, tendo como estudo de caso uma base de dados de uma operadora de planos de saúde suplementar contendo informações administrativas e de procedimentos em hospitais, laboratórios e consultórios relativas a beneficiários do estado de Santa Catarina, disponibilizados na forma de um banco de dados em *SQL Server* com o objetivo de encontrar padrões para auxiliar a previsão da propensão para determinados tipos de doenças em beneficiários com características semelhantes aos padrões encontrados. A qualidade do protótipo foi avaliada diante de dados devidamente preparados (pré-processamento) por especialistas da área, considerando o interesse nas regras encontradas pelo algoritmo.

A principal motivação desta dissertação é estudar o algoritmo cultural utilizando programação genética como componente populacional e avaliar a utilização das impressões gerais como um dos tipos de conhecimento dos algoritmos culturais. Outra motivação é encontrar conhecimento interessante e útil com foco em planos de saúde suplementar.

1.3 Objetivos

Pretende-se investigar a viabilidade da aplicação de um algoritmo cultural (fundamentado na combinação de programação genética e algoritmos culturais), no cotexto de planos de saúde suplementar, com o objetivo de descobrir conhecimento oculto e interessante (surpreendente, novo) em banco de dados, utilizando impressões gerais (conhecimento subjetivo) fornecidas pelo usuário como conhecimento de domínio no algoritmo cultural.

Os principais objetivos específicos neste trabalho são:

- Estabelecer uma nova abordagem para a tarefa de classificação em mineração de dados utilizando algoritmos culturais.
- Demonstrar o uso de impressões gerais do usuário em um algoritmo cultural.

- Encontrar padrões sobre dados de saúde suplementar.

1.4 Organização do Trabalho

Este trabalho está organizado da seguinte maneira:

O Capítulo 2 apresenta a revisão dos principais conceitos sobre descoberta de conhecimento, programação genética, algoritmos culturais e mineração de dados em planos de saúde. O Capítulo 5 apresenta todo o processo realizado para preparação dos dados que foram utilizados nesta dissertação. O Capítulo 3 apresenta como o algoritmo de programação genética foi desenvolvido e suas principais características para mineração de dados. O Capítulo 4 mostra o desenvolvimento do algoritmo cultural e suas principais características para ser utilizado em mineração de dados. No Capítulo 6 são apresentados os experimentos realizados e resultados. No Capítulo 7 são apresentadas as conclusões e os trabalhos futuros.

Capítulo

2

Revisão Bibliográfica

2.1 Considerações Iniciais

Este capítulo apresenta a revisão bibliográfica e a conceituação necessária para o entendimento desta dissertação. São descritos os principais conceitos sobre o processo de descoberta de conhecimento com ênfase nas etapas de pré-processamento e mineração de dados, além de conceitos sobre medidas de interesse no contexto de descoberta de conhecimento. Também são apresentados conceitos sobre programação genética e algoritmos culturais que são empregados na etapa de mineração de dados desta dissertação. Além desses conceitos são apresentados algumas abordagens encontradas na literatura sobre mineração de dados aplicado ao contexto de planos de saúde.

2.2 Descoberta de Conhecimento em Banco de Dados (DCBD)

O termo "Descoberta de Conhecimento em Banco de Dados"(DCBD) ou como é conhecido em inglês *KDD - Knowledge Discovery in Databases* - surgiu em um *workshop* em 1989 para enfatizar que "Conhecimento"é o produto final do processo de descoberta em banco de dados (Fayyad *et al.*, 1996b).

Frawley *et al.* (1992) definem o processo de descoberta de conhecimento como a extração não trivial de informações implícitas, previamente desconhecidas e potencialmente

úteis nos dados. Para Fayyad *et al.* (1996b) o termo DCBD é o processo não trivial de identificar padrões que sejam válidos, novos, potencialmente úteis e compreensíveis.

O termo "não trivial" significa que o processo de DCBD envolve algum mecanismo de busca ou inferência, ou seja, não é um cálculo simples de valores pré-definidos, como calcular o valor médio de um conjunto de números. Outra maneira de entender o terno "não trivial" é que um sistema de descoberta deve possuir algum grau de autonomia no processamento dos dados e avaliação dos resultados (Fayyad *et al.*, 1996b; Frawley *et al.*, 1992). Nessa definição, um conjunto de dados representa fatos enquanto que os padrões podem ser interpretados como uma expressão em alguma linguagem capaz de descrever um subconjunto de dados ou um modelo aplicável a esse subconjunto.

O processo de descoberta de conhecimento é dividido em várias etapas nas quais mineração de dados refere-se apenas à etapa de aplicação de algoritmos para extração de padrões em um conjunto de dados (Fayyad *et al.*, 1996b). Fayyad *et al.* (1996b) tem como ponto central justamente distinguir os termos "Descoberta de Conhecimento em Base de Dados"e "Mineração de Dados", os quais, muitas vezes são confundidos como sinônimos para identificar o processo de descoberta de conhecimento útil a partir de banco de dados. Além disso o processo de DCBD é interativo e iterativo, envolvendo repetições de etapas com muitas decisões que devem ser tomadas pelo usuário.

Os padrões descobertos pelo processo de DCBD devem ser novos, compreensíveis, úteis e interessantes (de acordo com a medida de interesse do usuário), com algum grau de certeza, ou seja, deverão trazer algum benefício novo que possa ser compreendido rapidamente pelo usuário para tomada de decisão. Esses padrões podem ser considerados conhecimento dependendo de sua natureza (Fayyad *et al.*, 1996b; Frawley *et al.*, 1992).

2.2.1 Etapas do processo de DCBD

Fayyad *et al.* (1996b) apresentam nove etapas do processo de descoberta de conhecimento. As etapas não estão necessariamente em ordem e podem ter múltiplas interações e repetições:

1. Desenvolver uma compreensão do domínio da aplicação, do conhecimento prévio relevante e identificar os objetivos do ponto de vista do cliente.

2. Criar um conjunto de dados alvo: selecionar um conjunto de dados ou focar em um subconjunto de variáveis ou dados em que a descoberta será realizada.

3. Limpar e pré-processar os dados: operações básicas tais como remoção de ruídos, se necessário, coleta da informação necessária para modelar ou estimar ruído, escolha

de estratégias para manipular campos de dados ausentes, formatação de dados de maneira a adequá-los à ferramenta de mineração.

4. Reduzir e projetar os dados: encontrar características úteis para representar os dados dependendo do objetivo da tarefa, visando a redução do número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados, bem como o enriquecimento semântico das informações.

5. Adequar o objetivo do processo de descoberta de conhecimento para uma tarefa de mineração de dados: tais como, sumarização, classificação, regressão, agrupamento, etc. As tarefas serão descritas em mais detalhes na seção 2.2.3.

6. Escolher o algoritmo de mineração de dados: escolher método(s) de seleção a ser utilizado para a busca de padrões de dados. Isso inclui decidir quais modelos e parâmetros podem ser adequados (por exemplo, modelos para dados categóricos são diferentes dos modelos de vetores sobre dados reais) e que correspondam a um método de mineração de dados específico com os critérios gerais do processo de DCBD (por exemplo, o usuário final pode estar mais interessado em compreender o modelo do que sua capacidade preditiva).

7. Mineração de Dados: buscar padrões em um formato de representação apropriado ao interesse do usuário, tais como: regras ou árvores de classificação. O usuário pode contribuir significativamente com o método de mineração de dados por meio da correta execução dos passos anteriores.

8. Interpretar os padrões encontrados possibilitando retornar aos passos 1 a 7 para mais interações. Esse passo pode também envolver visualização da extração de modelos/padrões ou visualização dos dados extraídos dos modelos.

9. Consolidação do conhecimento descoberto: incorporar o conhecimento em outros sistemas para apoiar ações ou simplesmente documentá-lo e submeter as partes interessadas. Inclui também verificar e resolver potenciais conflitos com crenças anteriores.

O processo de DCBD pode envolver significativa interação e pode conter iterações entre quaisquer etapas. O fluxo das etapas é ilustrado pela Figura 2.1. A maior parte do trabalho em DCBD é focado na etapa 7 - mineração de dados. Entretanto, há outras etapas importantes para o sucesso da aplicação do processo de DCBD na prática, como as etapas de pré-processamento que são importantes para tornar os dados mais apropriados para a etapa de mineração de dados (Fayyad *et al.*, 1996b; Tan *et al.*, 2006).

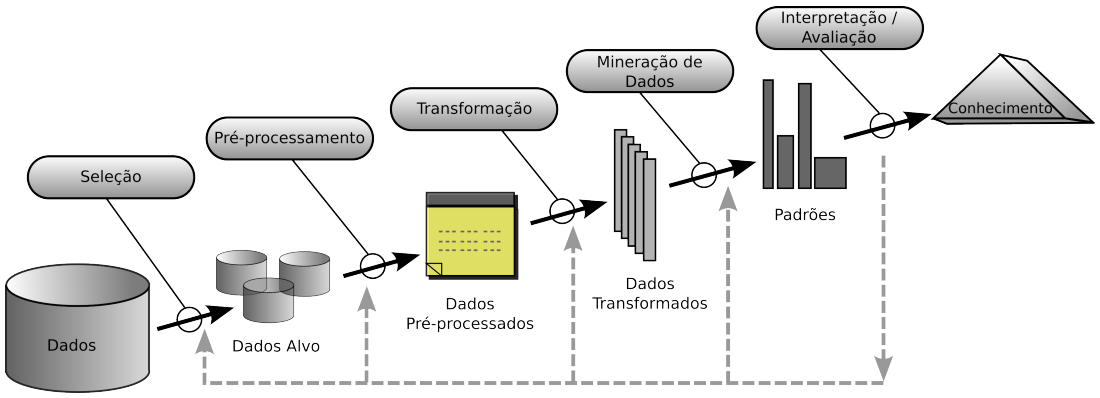

Figura 2.1: Uma visão dos passos do processo de KDD. (Fayyad *et al.*, 1996a)

2.2.2 Pré-Processamento

Normalmente, os dados disponíveis para análise não estão em um formato adequado para extração de conhecimento. Portanto alguns passos de pré-processamento devem ser aplicados para tornar os dados apropriados para a mineração de dados. O pré-processamento de dados é uma área abrangente e consiste em diferentes estratégias e técnicas que são inter-relacionadas, como a aplicação de métodos de tratamento, limpeza e redução do volume de dados (Rezende, 2003; Tan *et al.*, 2006).

Nesta dissertação utiliza-se o termo pré-processamento como toda a fase de preparação dos dados, que incluí desde o entendimento do domínio da aplicação até os dados prontos para servirem de entrada para os algoritmos de mineração de dados. Isso corresponde as quatro primeiras etapas do processo de DCBD, proposto por Fayyad *et al.* (1996b), com algumas adaptações. A seguir são apresentadas essas etapas com mais detalhes.

Entendimento do domínio da aplicação

O pré-processamento de dados em um processo de DCBD é frequentemente uma fase que envolve uma grande quantidade de conhecimento de domínio. Antes de ser iniciada qualquer operação de seleção, limpeza e transformação de dados é de fundamental importância que se tenha o conhecimento ou o entendimento dos dados, pois isso orienta as tomadas de decisões do que se investigar e o que será necessário ser realizado em termos de preparação de dados para mineração (Batista, 2003; Neves, 2003).

Segundo Neves (2003) o entendimento dos dados consiste em analisar os dados fornecidos pelos especialistas, entendendo do que se tratam as tabelas envolvidas, o significado, relevância, formato, tamanho e tipo de dado dos atributos; identificando

os atributos chaves; realizando levantamentos estatísticos e verificando a qualidade dos dados.

Batista (2003) propõe classificar as tarefas realizadas por métodos empregados na fase de pré-processamento em dois grupos: as tarefas fortemente dependentes do conhecimento de domínio e as fracamente dependentes do conhecimento de domínio.

As tarefas fortemente dependentes são as tarefas que somente podem ser realizadas com o uso de conhecimento específico do domínio. Por exemplo, em uma aplicação para concessão de crédito, a qual informa o valor emprestado, não pode assumir valores negativos. Ou ainda, caso existam informações a respeito do bem adquirido por meio de empréstimo, esse atributo pode possuir faixas de valores permitidos, as quais dependem do bem adquirido.

As tarefas fracamente dependentes do conhecimento de domínio, são tarefas que podem ser realizadas por métodos que extraem dos próprios dados as informações necessárias para tratar o problema de pré-processamento de dados. São exemplos de tarefas fracamente dependentes de domínio: o tratamento de valores desconhecidos e a identificação de valores extremos, entre outras.

Além do entendimento dos dados utilizados, o entendimento do domínio da aplicação como um todo pode ajudar os analistas na escolha do melhor conjunto de dados para se realizar a extração de padrões, saber quais valores são válidos para os atributos, os critérios de preferência entre os possíveis atributos, as restrições de relacionamento ou informações para geração de novos atributos (Rezende, 2003).

O sucesso do processo de DCBD depende, em parte, da participação dos especialistas do domínio da aplicação no fornecimento de conhecimento sobre o domínio e apoio aos analistas em sua tarefa de encontrar os padrões. Assim, antes do início das tarefas do processo é imprescindível a realização de um estudo a fim de adquirir um conhecimento inicial do domínio (Fayyad *et al.*, 1996b).

Seleção de um conjunto de dados alvo

Após o entendimento do domínio da aplicação é preciso selecionar o conjunto de dados dos quais serão extraídos os dados para mineração. Esses dados podem estar em um único ou vários bancos de dados, ou em um *Data Warehouse* que é um repositório de dados geralmente construído para dar suporte às pessoas que tomam decisões, tais como gerentes e diretores. A partir desses dados é preciso selecionar tabelas, atributos e instâncias das mesmas que estejam relacionadas aos objetivos do usuário, permitindo assim a geração de um novo conjunto de dados único e conciso, que poderá ser submetido às demais sub-fases

de pré-processamento bem como ao restante do processo de DCBD (Batista, 2003; Neves, 2003).

Para selecionar os dados é preciso descobrir onde estão armazenados nos bancos de dados, e geralmente é uma das fases mais trabalhosas de todo o processo de DCBD. Essa fase frequentemente envolve extrair dados de sistemas computacionais *legados*, ou seja, de sistemas antigos nos quais inexiste documentação a respeito do projeto de arquitetura do sistema. Dessa forma é necessário realizar muitas entrevistas e obter apoio dos especialistas de domínio para identificar onde determinada informação está registrada.

Os desafios encontrados para selecionar dados podem ser atenuados se a instituição em questão possuir um *Data Warehouse*. Essa tecnologia tem sido largamente utilizada, uma vez que os bancos transacionais não são considerados adequados para fornecer respostas para análises estratégicas. Segundo Batista (2003) os bancos de dados transacionais, sobretudo os de projetos mais antigos, frequentemente apresentam diversos problemas, tais como, problemas de falta de documentação de projeto, problemas com inconsistências e integridade de dados, entre outros. O *Data Warehouse* é periodicamente atualizado com dados de sistemas transacionais e/ou de fontes externas. Dados extraídos de diferentes bancos de dados são integrados e sua consistência é verificada, na medida do possível, antes de serem carregados.

Após selecionar a fonte de dados, tanto em bancos de dados convencionais quanto em *Data Warehouse* é necessário realizar a seleção e integração de tabelas realmente relevantes em relação aos objetivos a serem alcançados. Caso seja verificada a necessidade de manipular informações de várias tabelas de estruturas iguais e/ou diferentes é aconselhável integrar essas tabelas de modo a obter um conjunto único que será a base para o restante do processo de DCBD.

Limpeza de Dados

No mundo real os dados normalmente são incompletos, com ruídos e inconsistentes. As rotinas de limpeza de dados buscam preencher valores ausentes, suavizar ruídos, identificar valores extremos e corrigir inconsistências nos dados. Os problemas mais comuns encontrados nos dados são valores ausentes e ruído. Existem muitos métodos para resolver problemas de valores ausentes e ruído nos dados. Han e Kamber (2006) apresentaram seis métodos para lidar com ausência de valores e três métodos para ruído nos dados. Estes métodos serão descritos a seguir:

1. **Ignorar o registro**: É normalmente utilizado quando o valor da classe está faltando (assumindo que a tarefa de mineração seja de classificação). Este método não

é muito efetivo, a menos que o registro contenha diversos atributos com valores ausentes. É ruim especialmente quando a porcentagem de valores ausentes por atributo varia consideravelmente. Entretanto, diante de um conjunto de dados com uma quantidade muito grande de registros, é possível excluir aquelas que para um ou mais atributos apresentem valores ausentes, não representando uma perda de informação muito significativa.

2. **Completar valores ausentes manualmente:** Em geral, esta abordagem consome muito tempo, e pode não ser viável caso o conjunto de dados seja muito grande.

3. **Utilizar uma constante global para completar valores ausentes:** Substituir todos os valores ausentes de um atributo por uma mesma constante, tal como "desconhecido"ou "nenhum". Entretanto este método não é infalível, por exemplo, se os valores ausentes são substituídos por "desconhecido", o programa de mineração pode erroneamente julgar que eles formam um conceito interessante, pois todas elas possuem um valor comum.

4. **Utilizar a média do atributo para completar valores ausentes:** Por exemplo, admita a hipótese de que a média da renda anual dos clientes de uma loja seja de R$ 56.000. Então este valor deve ser utilizado para substituir os valores ausentes para o atributo renda.

5. **Utilizar a média do atributo para todos os exemplos pertencentes à mesma classe de um dado registro:** Por exemplo, se clientes são classificados de acordo com o risco de crédito, substitua o valor ausente de um determinado registro pelo valor com a renda média para os clientes da mesma categoria de risco de crédito.

6. **Utilizar o valor mais provável para completar valores ausentes:** Este valor deve ser determinado por regressão, ferramentas de inferência que utilizam formalismo Bayesiano ou árvore de indução. Por exemplo, utilizando outros atributos de clientes em seu conjunto de dados, pode-se construir uma árvore de decisão para predizer os valores ausentes para renda.

Valores ausentes geralmente ocorrem devido a erros humanos, pela não disponibilidade da informação no momento da entrada dos dados ou porque o atributo não é de preenchimento obrigatório. Nesses casos, a ausência de valores não implicam necessariamente em um erro nos dados. Por exemplo, quando um sistema de cadastro pede para um usuário

o número da licença de motorista. Os usuários que não tenham licença de motorista naturalmente deixarão esse campo vazio.

Dados ordenados para *preço*: 4, 8, 15, 21, 21, 24, 25, 28, 34

Partição dentro de intervalos (mesma frequência):

Intervalo 1: 4, 8, 15
Intervalo 2: 21, 21, 24
Intervalo 3: 25, 28, 34

Suavização por média de intervalos:

Intervalo 1: 9, 9, 9
Intervalo 2: 22, 22, 22
Intervalo 3: 29, 29, 29

Suavização por limites de intervalos:

Intervalo 1: 4, 4, 15
Intervalo 2: 21, 21, 24
Intervalo 3: 25, 25, 34

Figura 2.2: Métodos de Criação de intervalos. Fonte: (Han e Kamber, 2006)

Outro problema encontrado nos dados com frequência são os ruídos. Ruído é um erro aleatório ou uma variação na medição de uma variável. Dado um atributo numérico, tal como *preço*, como pode-se "suavizar"os valores e remover ruídos? A seguir são apresentadas algumas técnicas conforme (Han e Kamber, 2006):

1. **Criação de intervalos (*Binning*):** Os métodos de criação de intervalos buscam suavizar dados ordenados consultando a vizinhança do dado, isto é, os valores ao redor dele. Os valores ordenados são distribuídos dentro de um número de intervalos (*bins*). Pelo fato dos métodos de criação dos intervalos consultar a vizinhança de valores, eles realizam uma suavização local. A Figura 2.2 mostra algumas técnicas de criação de intervalos. Neste exemplo, dados para um atributo *preço* são primeiramente ordenados e então são particionados dentro de uma mesma frequência de intervalos de tamanho 3. No método de **suavização por média de intervalos**, cada valor pertencente ao intervalo é substituído pela média do intervalo. No método de **suavização por limites de intervalos**, os valores mínimos e máximo de um dado intervalo são identificados e cada valor do intervalo é substituído pelo limite mínimo ou máximo, respeitando a proximidade do valor original com o limite.

2. **Regressão:** Os dados podem ser suavizados ajustando-os por meio de uma função, tal como regressão: *Regressão linear* envolve encontrar a "melhor"maneira de ajustar dois atributos (ou variáveis), de modo que um atributo pode ser utilizado para predizer outro. *Regressão linear múltipla* é uma extensão da regressão linear, no qual mais que dois atributos estão envolvidos e os dados são ajustados por uma superfície multidimensional.

3. **Agrupamento:** Valores extremos podem ser detectados por agrupamento, no qual valores similares são organizados em grupos. Intuitivamente, valores que estão fora do conjunto podem ser considerados como valores extremos (Figura 2.3).

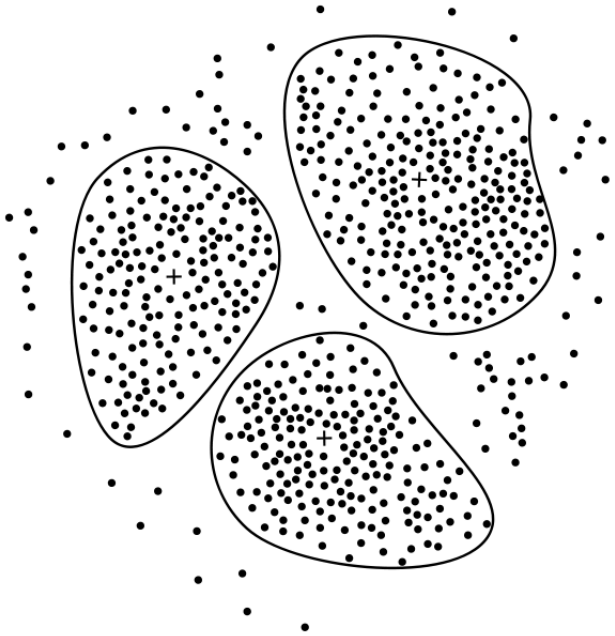

Figura 2.3: Um gráfico 2-D sobre dados de clientes e suas localizações em uma cidade, demonstrando três agrupamentos. Cada centroide é marcado com um "+", representando o ponto médio no espaço do grupo. Valores extremos podem ser detectados como valores fora do conjunto de grupos. Fonte: (Han e Kamber, 2006)

Muitos métodos de suavização também são métodos de redução de dados envolvendo discretização. Por exemplo, os métodos descritos acima reduzem o número de valores distintos por atributo. Conceitos de hierarquia são uma forma de discretização de dados que podem ser utilizados também para suavização dos dados. Um conceito hierárquico para *preço*, por exemplo, pode mapear os valores reais de *preço* em *barato*, *moderado* e *caro*, reduzindo desse modo o número de valores para serem manipulados pelo processo de mineração.

Pré-seleção e construção de atributos

Um conjunto de dados para análise pode conter centenas de atributos, muitos dos quais irrelevantes ou redundantes para a tarefa de mineração de dados. Por exemplo, se a tarefa for classificar clientes que são ou não potenciais compradores de um novo CD em uma loja quando fazem uma compra, atributos como telefone do cliente são provavelmente irrelevantes, diferente de atributos como *idade* e *gosto_musical* (Han e Kamber, 2006). Para remover esses atributos é feita uma pré-seleção de atributos ou seleção de um subconjunto de atributos.

Pré-seleção de atributos reduz o tamanho dos dados removendo atributos redundantes e irrelevantes, além de reduzir também a dimensionalidade dos dados. Segundo Riaño (1997), dimensionalidade refere-se ao número de atributos utilizados para descrever um conjunto de instâncias. Existem várias razões que justificam a redução do número de atributos no caso de excessiva dimensionalidade, como requisitos de tempo e espaço, simplicidade, relevância, redundância e acurácia, que serão discutidos no próximo parágrafo.

Requisitos de tempo e espaço: o custo de tempo e espaço dos algoritmos de indução estão diretamente relacionados ao número de atributos. *Simplicidade*: redução de atributos reflete na criação de estruturas menores que permitem um melhor entendimento do domínio, caso o erro não aumente significamente. *Relevância*: alguns atributos podem ser inúteis e portanto descartáveis. *Redundância*: alguns atributos podem refletir informações embutidas em outros atributos. *Acurácia (Accuracy)*: pré-processar o conjunto de propriedades pode melhorar a acurácia do modelo resultante.

A pré-seleção de atributos pode ser feita de maneira manual ou por meio de algoritmos. Na seleção manual o usuário pode auxiliar o analista de mineração de dados na escolha dos atributos que possam contribuir com a tarefa de mineração, assim como o analista pode também selecionar atributos com base em sua experiência sobre mineração de dados e seu conhecimento sobre o domínio do problema. Na seleção por algoritmos geralmente são utilizados heurísticas, pois para n atributos, há 2^n subconjuntos possíveis e portanto uma busca exaustiva pelo subconjunto ótimo de atributos pode ser extremamente custosa. Os melhores ou piores atributos são normalmente determinados utilizando testes de significância estatística, que assumem que os atributos são independentes entre eles. Muitas medidas de avaliação podem ser utilizadas, tal como ganho de informação, geralmente utilizado para construir árvores de decisão para classificação (Han e Kamber, 2006; Riaño, 1997).

Além disso, os atributos podem ser considerados inadequados para a mineração quando são fracamente ou indiretamente relevantes, condicionalmente relevantes ou medidos de

modo inapropriado. Entretanto, esses atributos podem ser convenientemente combinados gerando novos atributos que podem mostrar-se altamente representativos para descrição de um conceito (Batista, 2003). Para combinar esse atributos pode-se utilizar um processo de construção de novos atributos conhecido como *construção de atributos* ou *indução construtiva*[1] (Bloedorn e Michalsi, 1998).

De maneira bastante ampla, o processo de indução construtiva pode ser dividido em duas abordagens: a automática e a guiada pelo usuário. A *indução construtiva automática* consiste em um processo de construção de atributos guiada automaticamente pelo método de construção. Geralmente, os atributos construídos são avaliados em relação aos dados, e podem ser descartados ou integrados ao conjunto de dados. A *indução construtiva guiada pelo usuário* utiliza o conhecimento do usuário ou do especialista do domínio para guiar a composição dos atributos (Batista, 2003).

Redução e Projeção de Dados

Reduzir e projetar os dados significa encontrar características úteis para representar os dados dependendo do objetivo da tarefa, visando a redução do número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados, mantendo a integridade original dos dados bem como o enriquecimento semântico das informações. A redução do conjunto de dados deve deixar a mineração de dados mais eficiente e ainda assim produzir os mesmos resultados (ou quase os mesmos) (Fayyad *et al.*, 1996b; Han e Kamber, 2006).

Existem várias técnicas e estratégias para redução de dados, a seguir apresenta-se cinco técnicas conforme Han e Kamber (2006):

1. **Agregação em um cubo de dados**, em que operações de agregação são aplicadas sobre os dados para construção de um cubo de dados.

2. **Seleção de um subconjunto de atributos**, no qual atributos irrelevantes, fracamente relevantes ou redundantes devem ser detectados e removidos.

3. **Redução de dimensionalidade**, em que mecanismos de codificação são utilizados para reduzir o tamanho dos dados.

4. **Redução de numerosidade**, no qual os dados são substituídos ou estimados por alternativas, representação de dados menores tais como modelos paramétricos (que precisam armazenar apenas os parâmetros do modelo em vez dos dados reais)

[1] Os termos *aprendizado construtivo* ou *transformação de atributos* também são encontrados na literatura

ou métodos não-paramétricos tais como agrupamento, amostragem e o uso de histogramas.

5. **Discretização e geração de conceitos hierárquicos**, no qual dados brutos dos atributos são substituídos por níveis conceituais mais altos. Discretização dos dados é uma maneira de redução de numerosidade que é muito útil para geração automática de conceitos hierárquicos. Discretização e geração de conceito hierárquico são poderosas ferramentas para mineração de dados, na medida em que permitem mineração de dados em múltiplos níveis de abstração.

Nesta dissertação utilizou-se histogramas. Histogramas utilizam intervalos para aproximar as distribuições de dados e são uma maneira popular de redução dos dados. Um histograma para um atributo A pode particionar os dados de A em subconjuntos disjuntos ou *buckets*. Para reduzir ainda mais os dados, é comum para cada subconjunto utilizar uma faixa de valores contínuos para um dado atributo.

2.2.3 Mineração de Dados

A etapa de Mineração de Dados (MD) envolve decidir quais algoritmos serão aplicados aos dados para encontrar algum conhecimento interessante.

Para encontrar respostas ou extrair conhecimento relevante, existem diversas técnicas de MD disponíveis na literatura (Chen *et al.*, 1996), tais como:

- Métodos estatísticos;
- Redes neurais;
- Árvores de decisão;
- Algoritmos heurísticos.

O uso da estatística descritiva e inferencial é muito difundido em projetos de MD. A utilização de técnicas estatísticas requer o uso de dados quantitativos. Dados qualitativos devem ser traduzidos em valores numéricos. Os testes estatísticos podem ser usados para a comparação de valores de amostras de grupos diferentes. Nos casos mais simples, a estatística descritiva é utilizada para obter uma visão geral das características dos dados analisados. A estatística descritiva inclui medidas como a média, mediana (valor que divide a distribuição pela metade), moda (valor mais comum), desvio-padrão (medida de variância), intervalos de confiança (maior e menor valores) e distribuição dos dados.

Para a utilização de testes estatísticos, algumas condições devem ser satisfeitas, como a utilização de valores numéricos quantitativos e a existência de uma hipótese que deve ser posta em teste.

As árvores de decisão podem ser utilizadas para a classificação de determinados tipos de dados (Breiman *et al.*, 1984). Seu nome é originado da característica de sua estrutura, que é formada por ramos e nós. Em cada nó, uma decisão é tomada ou a ocorrência de uma probabilidade é avaliada. Dependendo da direção escolhida, um outro ramo será seguido levando ao próximo nó de decisão.

Supondo, por exemplo, que o objetivo é saber se vai ou não ter jogo baseando-se na previsão do tempo. Nesse caso os dados analisados contêm características do tempo, como: vendaval, temperatura, sol, chuva e nublado. Os algoritmos utilizados para subdividir os dados procuram os atributos que conseguem a máxima separação das variáveis. Nesse exemplo, a primeira divisão considera o atributo previsão: sol, chuva ou nublado. Essa é a variável que divide igualmente o conjunto de dados. Teoricamente qualquer atributo poderia ser escolhido, mas os algoritmos escolhem aqueles que segregam ao máximo o número de registros, trazendo significância para a árvore de decisão.

Após a construção da árvore de decisão, como na Figura 2.4, inferências sobre os dados podem ser feitas. Por exemplo:

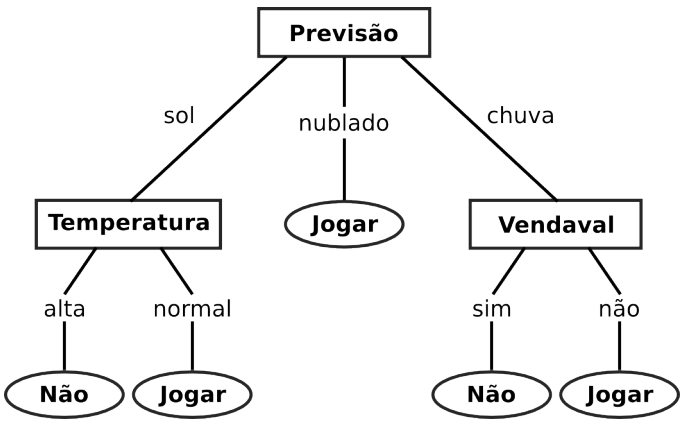

Figura 2.4: Estrutura de uma árvore de decisão. Fonte: (Coelho, 2004)

- Se a previsão for sol **E**
- Se a temperatura for alta **ENTÃO**
- Possui uma alta probabilidade de não haver jogo

Dessa forma, as árvores de decisão são úteis quando o principal objetivo é determinar categorias para classificação ou previsão. Por outro lado, são pouco úteis quando a aplicação consiste em previsões específicas sobre os valores de variáveis quantitativas.

Outra abordagem muito utilizada na atualidade são os algoritmos heurísticos, que inclui diversas técnicas, tais como: Algoritmo das Abelhas (Pham *et al.*, 2007), Evolução Diferencial (Storn e Price, 1997), Anelamento Simulado (Kirkpatrick *et al.*, 1983), Busca Harmônica (Geem, 2006), *Intelligent Water Drops* (Shah-Hosseini, 2009), *Beam search* (Furcy e Koenig, 2005), *Gravitational Search Algorithm* (Abarghouei *et al.*, 2009), algoritmos genéticos (Romão *et al.*, 2004), programação genética (Koza, 1992) e *Ant Colony Optimization* (Colorni *et al.*, 1992). Nesta dissertação o foco é a técnica de programação genética que é discutida posteriormente, pois está associada à otimização e é de extrema importância para esta dissertação.

Essas técnicas podem ser aplicadas a diversas tarefas de MD, tais como: agrupamento, associação, classificação, regressão e modelagem de dependência, das quais as principais tarefas são as de agrupamento, associação e classificação.

A tarefa de agrupamento (*clustering*) busca fornecer uma descrição dos dados (Maimon e Rokach, 2005). Segundo Romão (2002) agrupamento é a tarefa em que o algoritmo deve descobrir a classe de cada exemplo baseado na sua similaridade com os demais exemplos do conjunto de dados, dividindo os exemplos em grupos. Veja o exemplo da Figura 2.3. Além disso, um algoritmo de agrupamento deve ser capaz de maximizar a semelhança entre os elementos de um mesmo grupo e minimizar as semelhanças entre exemplos pertencentes a grupos diferentes.

A tarefa de associação tem como principal objetivo encontrar, a partir de um conjunto de exemplos, um conjunto de regras de associação, ou seja, descobrir quais atributos aparecem frequentemente associados nesses exemplos. Esse tipo de tarefa é normalmente aplicada em um tipo especial de dados, como grandes quantidades de dados relativos a compras de clientes em supermercados. Esses tipos de dados são comumente conhecidos como "transações de cesta de compras"(Han e Kamber, 2006).

A tarefa de classificação é uma tarefa extremamente importante para este trabalho e será descrita em maiores detalhes na próxima seção.

Tarefa de Classificação

A classificação é um importante problema extensivamente estudado em diversas áreas de pesquisa, tais como reconhecimento estatístico de padrões, aprendizagem de máquina e mineração de dados (Farrar *et al.*, 1999; Fayyad *et al.*, 1996a; Witten e Frank, 2005). Em

MD é provavelmente a tarefa mais utilizada e a mais importante para a presente pesquisa. Essa tarefa consiste em classificar uma instância (exemplo ou registo) como pertencente a uma determinada classe dentre várias classes previamente definidas (Carvalho, 2005; Coelho, 2004).

Os dados utilizados para resolver esse tipo de tarefa consistem em um conjunto de atributos denominados previsores e um atributo denominado meta, que define a classe a que cada registro pertence. O principal objetivo da construção de um classificador é descobrir um relacionamento entre os atributos previsores e o atributo meta usando registros cuja classe é conhecida. Esse relacionamento poderá ser utilizado posteriormente para prever a classe de um registro, cuja classe é desconhecida, com uma determinada probabilidade de acerto, ou entender a relação existente entre os atributos previsores e a classe (Breiman *et al.*, 1984; Hand, 1997).

Para entender a relação existente entre os atributos previsores e a classe é exigido do classificador que ele não apenas classifique, mas também explique o conhecimento extraído da base de dados de maneira compreensível.

A fim de contribuir com a compreensibilidade do conhecimento descoberto (relação entre atributos e as classes), pode-se construir um classificador baseado em regras, que é uma técnica para classificar exemplos utilizando um conjunto de regras do tipo "SE...ENTÃO". Essas regras também são conhecidas como regras de decisão e constituem uma forma de representação simbólica de regras de previsão com o seguinte formato (Romão, 2002; Tan *et al.*, 2006):

$$SE < antecedente > ENTAO < consequente >$$

O antecedente é formado por uma expressão lógica envolvendo pares atributo-valor do domínio da aplicação existentes nos bancos de dados. O consequente é formado por uma expressão que indica a previsão de algum valor para o atributo meta, obtido em função dos valores encontrados nos atributos que compõem o antecedente.

Um atributo é chamado de *atributo meta* quando seu valor ou classe for previsto, isto é, estiver do lado do consequente da regra.

Portanto, o objetivo da tarefa de classificação pode ser a de descobrir regras capazes de prever o valor de um atributo meta a partir dos valores de atributos previsores.

Avaliação do conhecimento descoberto

Independente do tipo de algoritmo utilizado na tarefa de classificação, o conhecimento descoberto deve ser avaliado. Medir a qualidade do conhecimento extraído por meio de

um algoritmo de mineração de dados não é uma tarefa simples, uma vez que múltiplos objetivos devem ser considerados durante a avaliação, sendo alguns deles muito subjetivos (Freitas, 2002b).

Existem diversas medidas de avaliação dentre as quais pode-se citar:

- Taxa de Acerto
- Grau de Interesse
- Padrões
- Credibilidade
- Exceção confiável
- Peculiaridade

Para classificadores baseados em regras, um critério muito conhecido é a taxa de acerto. Romão (2002) apresenta uma maneria simples de medir a taxa de acerto, como mostra a Equação 2.1:

$$Taxa\ de\ acerto = \frac{Num.\ de\ instâncias\ dos\ dados\ classificados\ corretamente}{Num.\ de\ instâncias\ dos\ dados\ aplicáveis} \quad (2.1)$$

Tanto a Equação 2.1 como vários outros métodos de avaliação de regras podem ser calculadas com auxílio de uma *matriz de confusão*. A matriz de confusão é um método útil para analisar o quão bem seu classificador pode reconhecer exemplos de diferentes classes. Uma matriz de confusão de duas classes é exibida pela Figura 2.5

		Classe real do exemplo	
		C_1	C_2
Classe prevista	C_1	VP	FP
	C_2	FN	VN

Figura 2.5: Uma matriz de confusão para exemplos positivos e negativos

O rótulo de cada quadrante da matriz possui o seguinte significado:

- VP = Verdadeiros positivos = Exemplos positivos que foram corretamente rotulados pelo classificador.

- VN = Verdadeiros negativos = Exemplos negativos que foram corretamente rotulados pelo classificador.
- FP = Falso positivos = Exemplos negativos que foram incorretamente rotulados como positivos.
- FN = Falso negativos = Exemplos positivos que foram incorretamente rotulados como negativos.

Note que a taxa de acerto para a classe C1 calculada pela Equação 2.1 pode ser definida também utilizando-se os termos da matriz de confusão como:

$$Taxa\ de\ acerto = \frac{VP}{VP + FP} \tag{2.2}$$

Entretanto a matriz de confusão é utilizada para avaliar o conhecimento descoberto sem considerar o interesse do usuário sobre o conhecimento descoberto. Para levar em consideração os interesses do usuário existem as medidas de interesse. Essas medidas serão descritas a seguir.

Medidas de Interesse

O processo de DCBD pode gerar conhecimento exato, compreensível e de forma eficiente, entretanto esse conhecimento pode não ser de interesse do usuário. Muitos métodos para descoberta de conhecimento encontram um grande número de regras, mas a maioria dessas regras não representa informação relevante para o usuário. Além disso, o grande número de regras dificulta a análise pelo usuário para encontrar regras relevantes. É recomendável utilizar algum método que meça o **grau de interesse** do conhecimento descoberto.

Existem vários métodos propostos na literatura para avaliar o grau de interesse das regras descobertas. Em geral são organizadas em dois grupos: *data-driven* (métodos objetivos) e *user-driven* (métodos subjetivos) (Freitas, 1998; Silberschatz e Tuzhilin, 1996).

Os métodos *data-driven* tentam estimar o quanto as regras podem ser interessantes ao usuário de forma mais automática e indireta, sem exigir que esse especifique suas crenças ou conhecimento prévio. Esses métodos têm a desvantagem de serem uma estimativa indireta do quão surpreendente serão as regras para o usuário, ignorando suas crenças ou conhecimento prévio. Porém, também existem algumas vantagens, por exemplo uma maior independência do domínio da aplicação e um grau maior de automatização, liberando o usuário da tarefa de fornecer suas crenças ou conhecimento prévio (Carvalho, 2005).

Um exemplo do método *data-driven* de avaliação de regras é a confiança e o suporte das regras, obtidos pela tarefa de associação. Para a tarefa de classificação, pode-se empregar uma medida baseada na teoria da informação. Mais detalhes sobre essa medida de interesse pode ser encontrada em (Freitas, 1998; Romão, 2002).

A ideia básica dos métodos *user-driven* é que o usuário especifica suas crenças (impressões gerais) ou conhecimento prévio sobre o domínio da aplicação e o sistema utilize esse tipo de informação para selecionar as regras interessantes. Uma regra é considerada interessante se ela representar alguma novidade com relação às crenças ou conhecimento do usuário. Esses métodos têm a vantagem de considerar diretamente as crenças do usuário, porém têm a desvantagem de serem fortemente dependentes do conhecimento do domínio da aplicação e serem menos automáticas do que as medidas *data-driven*, exigindo uma participação intensiva do usuário na tarefa de tornar explícitas as suas crenças e avaliações.

Esta dissertação utiliza um método *user-driven* baseado em impressões gerais proposto por Romão (2002) para analisar regras de classificação. Impressões Gerais (IG's) são um tipo específico de conhecimento prévio que o usuário tem a respeito do domínio da aplicação.

Muitas vezes o usuário não sabe detalhes a respeito do domínio, como alguns sistemas especialistas exigem, mas normalmente ele sempre tem algumas IG's. Por exemplo, um analista de plano de saúde pode pensar que beneficiários com mais de 60 anos e que não realizam exercícios físicos possuem grande chance de desenvolver diabetes.

Romão (2002) adaptou uma técnica proposta por Liu e Hsu (1996) em que o conhecimento prévio do usuário é representado por um conjunto de impressões gerais, denominado IG, e cada conjunto de regras é denominado R. A Figura 2.6 ilustra como é feita a avaliação das regras em que se interessa saber o grau de similaridade, e a diferença entre cada regra em R comparada com cada impressão geral existente no conjunto IG. As regras IG e R devem ter a mesma sintaxe e semântica. Além disso a técnica de Romão (2002) utiliza lógica difusa, buscando facilitar as comparações, em que são definidas variáveis linguísticas para atributos com valores contínuos.

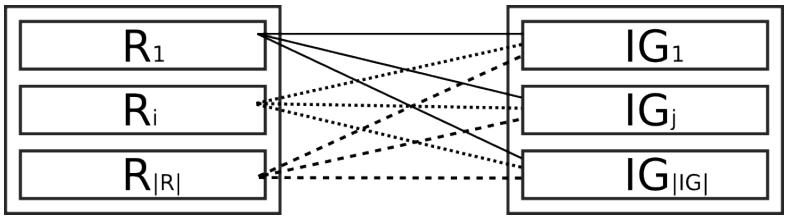

Figura 2.6: Comparação da regra com IG. Fonte: (Romão, 2002)

Métodos de Validação

Para se obter uma estimativa confiável das medidas de avaliação de conhecimento citadas na subseção anterior, existem vários métodos, os mais comumente utilizados são *Holdout*, Sub-Amostragem Aleatória, Validação Cruzada e *Bootstrap*. Esses métodos têm como objetivo avaliar a precisão baseado em amostras aleatórias de partições dos dados. Apresenta-se aqui a *holdout* e validação cruzada.

No método *holdout*, os dados originais com exemplos rotulados são particionados em dois conjuntos disjuntos, chamados de conjunto de treinamento e conjunto de teste, como ilustrado na Figura 2.7. Um modelo de classificação é então induzido a partir do conjunto de treinamento e seu desempenho é avaliado utilizando-se o conjunto de teste. A proporção dos conjuntos são geralmente dois terços para treinamento e um terço para teste, mas essa proporção pode ser diferente (Tan *et al.*, 2006).

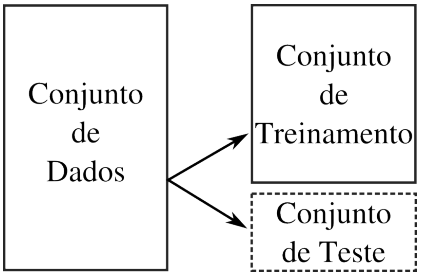

Figura 2.7: Separação dos Dados

Na validação cruzada, os dados são particionados em k subconjuntos mutuamente exclusivos, $D_1, D_2, ..., D_k$, cada um com aproximadamente o mesmo tamanho. O treinamento e o teste são realizados k vezes. Na iteração i, a partição D_i é reservada para teste, as demais partições são coletivamente utilizadas para treinar o modelo. Ou seja, na primeira iteração, os subconjuntos $D_2, ..., D_k$ são coletivamente utilizados como conjunto de treinamento para obter o primeiro modelo, que é então testado sobre D_1; na segunda iteração, o modelo é treinado utilizando-se os subconjuntos $D_1, D_3, ..., D_k$ e testado sobre D_2; e assim sucessivamente. A Figura 2.8 ilustra o funcionamento da validação cruzada para $k = 4$. Diferentemente do *holdout*, na validação cruzada cada exemplo é utilizado o mesmo número de vezes para treinamento e uma vez para teste. Para classificação, a precisão estimada é sobre o número total de exemplos de teste classificados corretamente para as k iterações, dividido pelo número total de exemplos dos dados originais (Han e Kamber, 2006).

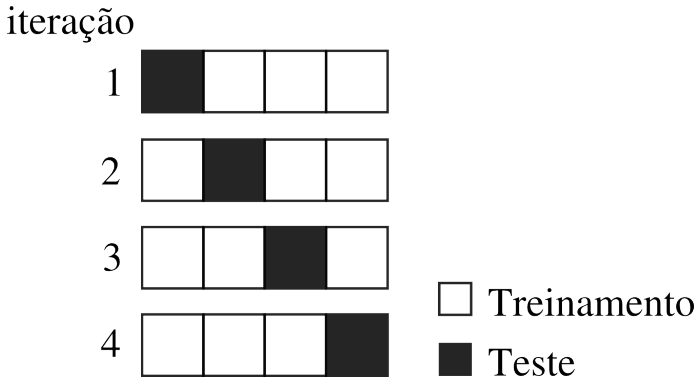

Figura 2.8: Validação Cruzada com 4 partições

2.3 Programação Genética

Programação Genética (PG) é um paradigma que pertence à família dos Algoritmos Evolucionários (AEs). Os algoritmos evolucionários são métodos de busca e otimização baseados na abstração da teoria de evolução por seleção natural de Darwin, no qual sobrevivem os indivíduos que possuem maior capacidade de se adaptarem ao meio ambiente. A carga genética desses indivíduos é passada aos seus descendentes por meio de cruzamento, e sofre variações (mutação) que podem ser benéficas ou não para sobrevivência. Durante o processo evolutivo, variações úteis ou favoráveis dos indivíduos tem uma probabilidade maior de serem preservadas, enquanto que variações prejudiciais tem uma probabilidade maior de serem descartadas.

Programação Genética foi proposta por Koza (1992) e surgiu a partir da extensão das ideias dos algoritmos genéticos (AG), um outro tipo particular de algoritmo evolucionário, proposto por Holland (1992). No (sub)paradigma de algoritmos genéticos, assim como em programação genética, as possíveis soluções dos problemas são tratadas como indivíduos de uma população que irá evoluir ao longo das gerações. Essas populações de indivíduos sofrerão um processo de evolução utilizando-se de operadores de reprodução, cruzamento e mutação de modo a gerar uma nova população constituída por indivíduos melhores. O processo é guiado por uma função de aptidão (*fitness*) que avalia a capacidade de adaptação de um indivíduo ao ambiente, ou seja, o quanto o indivíduo está perto da solução do problema. Os indivíduos que possuírem maior capacidade de adaptação terão maiores chances de sobreviver.

Entretanto há algumas diferenças importantes entre estes dois (sub)paradigmas de AEs. Um das diferenças dizem respeito à questão da representação do indivíduo e correspondentes operadores genéticos. A maioria dos algoritmos de PG representam seus

indivíduos como árvores. Nesse caso, um indivíduo (solução candidata) é representado por uma árvore, na qual os nós internos são funções e os nós terminais são variáveis ou constantes do problema. Indiscutivelmente, a mais importante distinção entre a representação de um indivíduo em PG e AG é que em PG podem conter não somente valores de variáveis (dados), mas também funções, diferentemente de indivíduos em AG, que normalmente possuem somente dados. Assim, um indivíduo em PG pode ser chamado de um "programa", embora algumas vezes em um sentido amplo da palavra (Freitas, 2002b).

Ao longo dos anos a programação genética vem sendo aplicada em diversas áreas do conhecimento, como Engenharia de *Software* (Forrest *et al.*, 2009), Circuitos Digitais (Carneiro *et al.*, 2008), Previsão de Séries Temporais (De Souza, 2006), Mineração de Dados (Bojarczuk *et al.*, 2004) entre outras.

2.3.1 Representação dos programas

A mais importante distinção entre AG e PG é que em PG um indivíduo pode conter não somente valores de atributos, mas também funções, enquanto que os indivíduos em algoritmos genéticos tipicamente contêm somente atributos. Um indivíduo em PG pode ser chamado de um "programa"e deve corresponder a uma receita genérica para resolver um dado tipo de problema, enquanto que um indivíduo em AG tipicamente corresponde a uma solução particular para um determinado tipo de problema.

Os "programas"são formados pela combinação de funções e terminais, adequados ao domínio do problema. Tem-se então dois conjuntos: o conjunto de funções F e o conjunto de terminais T. O conjunto F pode conter operadores aritméticos (*+, -, *, etc.*), funções matemáticas (*seno, log, etc.*), operadores lógicos (*e, ou, etc.*), entre outros. O conjunto T é composto pelas variáveis e constantes e fornece um valor para o sistema enquanto o conjunto de funções processa os valores no sistema. Juntos, os conjuntos de funções e terminais representam os nós.

Pode-se citar, como exemplo, o conjunto F dos operadores aritméticos de aridade[2] dois e o conjunto T de terminais, da seguinte forma:

$$F = \{+, -, *, /\} \quad T = \{A, B, C\} \tag{2.3}$$

A combinação desses dois conjuntos pode formar um indivíduo como: $A + (B * C)$ que pode ser representado em forma de árvore como na Figura 2.9.

[2]Número de argumentos de uma função

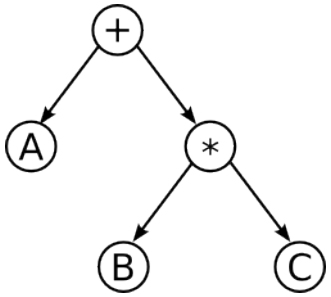

Figura 2.9: Representação de um indivíduo (programa) em PG

2.3.2 Fechamento e Suficiência

Em geral, o conjunto de funções de um algoritmo de programação genética deve satisfazer três propriedades: Suficiência (*sufficiency*), Parcimônia (*parsimony*) e Fechamento (*closure*) (Banzhaf *et al.*, 1998). Para garantir a convergência para uma solução, Banzhaf *et al.* (1998) definiu a propriedade de Suficiência como o poder representativo do conjunto de funções e terminais, ou seja, os conjuntos devem ser bons o suficiente para representar uma solução do problema que se deseja resolver e Parcimônia que considera uma abordagem parcimoniosa para escolha de um conjunto de funções mais inteligente (Banzhaf *et al.*, 1998).

Fechamento significa que uma função deve ser capaz de aceitar, como entrada, qualquer saída produzida por qualquer função ou terminal. Na prática há diversas situações nas quais PG deve lidar com variáveis de diferentes tipos de dados, o que torna difícil satisfazer a propriedade de fechamento. Essa imposição garante que qualquer árvore gerada possa ser avaliada corretamente. Um exemplo representativo do problema de fechamento é a operação de divisão: matematicamente não é possível dividir um valor por zero.

2.3.3 População Inicial

O primeiro passo na execução em PG é inicializar a população. Isso significa criar uma variedade de estruturas de programas para posterior evolução. Essa variedade de estruturas ou conjunto de indivíduos é gerada aleatoriamente a partir dos conjuntos de funções F e de terminais T. O processo é realizado escolhendo-se funções e terminais para a composição da árvore, até que se atinja a profundidade máxima definida para a árvore. Existem vários métodos para geração aleatórias de árvores para PG, como os métodos *Full*, *Grow*, *Ramped half-and-half* (Koza, 1992), *Probabilistic Tree-Creation* (PTC) 1 e 2 (Luke, 2000), descritos a seguir:

Full

O método *full* de geração aleatória de indivíduos cria árvores completas. Para isso somente escolhe apenas funções até o nó chegar a profundidade máxima da árvore. Então ele escolhe somente terminais.

Grow

O método *grow* gera árvores de formas irregulares. A escolha dos nós é feita de forma aleatória entre funções e terminais. A profundidade de um nó qualquer da árvore até o nó raiz está restrita à profundidade máxima definida para um indivíduo.

Ramped half-and-half

O método *Ramped half-and-half* é uma combinação dos métodos *grow* e *full*, isto é, cada um dos métodos é escolhido 50% das vezes. O método também envolve a criação de um número igual de árvores utilizando um parâmetro de profundidade entre 2 e a profundidade máxima especificada. Por exemplo, se a profundidade máxima especificada for 6, então 20% das árvores serão de profundidade 2, 20% das árvores serão de profundidade 3, e assim por diante até a profundidade máxima.

Probabilistic Tree-Creation (PTC1) e (PTC2)

O método *Probabilistic Tree-Creation 1* é uma modificação do método *grow* e permite ao usuário fornecer as probabilidades das funções que podem aparecer na árvore, mais uma expectativa para o tamanho da árvore, e garante que, em média, as árvores serão do tamanho esperado.

O algoritmo PTC1 é dado como segue: durante a geração da árvore, o algoritmo alternativamente escolhe o nó filho a partir do conjunto de funções F ou terminais T. Para cada função f em F, é indicada uma probabilidade q_n de escolha sobre o conjunto F para quando o algoritmo precisar de uma função. Similarmente, para cada terminal t em T, é indicada uma probabilidade q_t de escolha sobre o conjunto T para quando o algoritmo precisar de um terminal. É indicado também a profundidade máxima D, embora esse limite seja utilizado somente para forçar uma absoluta, e preferencialmente grande restrição de memória. E finalmente, é indicado um tamanho esperado para a árvore E_{tree}.

Antes de gerar todas as árvores, o algoritmo calcula p, que é a probabilidade de escolher uma função ou um terminal a fim de produzir uma árvore com um tamanho esperado E_{tree}:

$$p = \frac{1 - \frac{1}{E_{tree}}}{\sum_{n \in N} q_n b_n} \tag{2.4}$$

onde b_n é a aridade da função n. Esse cálculo precisa ser feito somente uma vez. Em seguida o algoritmo prossegue para construir a árvore. A Figura 2.10 mostra o pseudo-código do algoritmo.

```
Dados:
   Profundidade máxima D
   Conjunto de funções F e de terminais T
   A probabilidade p de escolher uma função
   As probabilidades q_t e q_f para cada t ∈ T e f ∈ F
Faça:
   Árvore = PTC1(0)

PTC1(profundidade d)
   Retornar: uma árvore de profundidade máxima ≤ D - d

   Se d = D
        Retornar um terminal de T (baseado em q_t)
   Senão
        Com probabilidade p, Se uma função deve ser escolhida
             Escolha aleatoriamente f ∈ F baseado em q_f
             Para cada argumento a de f
                  Preencha a com PTC1(d+1)
             Retornar f com todos os argumentos preenchidos
        Senão
             Retornar aleatoriamente um terminal de T baseado em q_f
```

Figura 2.10: Algoritmo PTC1

Uma variante desse método (PTC2) utiliza um tamanho máximo S e uma distribuição de probabilidade $w_1, w_2, ..., w_s$ para cada árvore de tamanho 1 a S e, além do controle sobre o tamanho esperado da árvore, ele permite também o controle sobre a distribuição desses tamanhos.

2.3.4 Função de Aptidão

Segundo (Koza, 1992) a função de aptidão (*fitness*) é a força motriz tanto em programação genética quanto em algoritmos genéticos. Na natureza, a aptidão de um indivíduo é a probabilidade de sobrevivência desse indivíduo até a idade de reprodução. Em programação genética deve-se medir a aptidão de acordo com o problema sendo resolvido e então utilizar essa medida para controlar a aplicação de operações que modificam as estruturas da população.

Para medir a aptidão é necessário definir uma função de aptidão, que é feita de acordo com o domínio do problema. Boas escolhas são essenciais para se obterem bons resultados, já que a função de aptidão é que direciona o algoritmo de PG na busca por boas soluções.

A abordagem mais comum para a medição de aptidão é criar uma medida explícita para cada indivíduo da população. Essa abordagem é utilizada na maioria das aplicações convencionais de AG. A cada indivíduo da população é atribuído um valor escalar de aptidão que é calculado por algum procedimento de avaliação bem definido (função de aptidão). Essa aptidão é que permite dizer se um indivíduo conseguirá existir, sobreviver e ter sucesso em se reproduzir em uma população (como ocorre na natureza) (Koza, 1992).

2.3.5 Métodos de Seleção

Depois que a qualidade de um, indivíduo foi determinada pela aplicação da função de aptidão, é preciso decidir se serão aplicados os operadores genéticos sobre esse indivíduo e ele se o mesmo será mantido na população ou se será substituído. Esta tarefa é chamada de seleção e é atribuída a um operador especial, o operador de seleção (Banzhaf *et al.*, 1998).

Há diversos tipos de operadores de seleção tais como: seleção proporcional (*proportional selection*), truncamento (*truncation selection*), ranqueamento (*ranking selection*), e seleção por torneio. Segundo Banzhaf *et al.* (1998), a decisão sobre qual método de seleção aplicar é uma das decisões mais importantes a serem tomadas durante a execução em PG. A seleção é responsável pela velocidade da evolução e é frequentemente citado como o culpado em casos em que a convergência é prematura. Para fins desta dissertação apresenta-se os métodos de seleção proporcional e por torneio.

Seleção Proporcional

Esse método foi apresentado por Holland (1992) para ser aplicado sobre AG. Nessa abordagem, os indivíduos são copiados para a próxima geração de acordo com probabilidades

proporcionais ao seu valor de função da avaliação. A implementação desse método é feito por meio de um mecanismo de roleta, na qual a roleta é dividida em N partes em que N é o número de indivíduos da população. O tamanho de cada uma das partes da roleta é proporcional à aptidão do indivíduo que a parte representa. A roleta é então girada N vezes e a cada uma delas o indivíduo indicado pelo ponteiro é selecionado e inserido na nova população.

O método da roleta pode levar a um grande número de cópias de indivíduos bons, diminuindo assim a diversidade da população. Isso pode induzir uma convergência prematura do algoritmo a uma solução não desejada (mínimo/máximo local).

Seleção por Torneio

A seleção por torneio não é baseado na competição sobre toda a população, mas sobre um subconjunto da população. Um número k de indivíduos, chamado de tamanho do torneio, são selecionados aleatoriamente da população, e então as aptidões desses indivíduos são comparadas. O indivíduo com a maior aptidão é selecionado para reprodução. Esse processo é repetido uma quantidade de vezes suficiente para formar uma nova população.

Segundo Pappa (2002) o valor de k é defino pelo usuário. Quanto maior o valor de k, maior a pressão seletiva, ou seja, maior a velocidade com que os indivíduos mais fortes dominam a população, causando a extinção dos mais fracos.

2.3.6 Operadores Genéticos

Geralmente, após a inicialização da população, a aptidão dos indivíduos é baixa. A evolução é realizada transformando a população inicial por meio dos operadores genéticos. Em termos de aprendizagem de máquina, esses são operadores de pesquisa. Embora existam muitos tipos de operadores genéticos, os três principais segundo Banzhaf *et al.* (1998) são:

- Cruzamento;
- Mutação; e
- Reprodução

A seguir apresenta-se o funcionamento de cada um desses operadores.

Cruzamento

O operador de cruzamento é responsável por combinar material genético entre dois indivíduos. Para isso troca-se parte de um desses indivíduos por uma parte do outro indivíduo. Em PG o cruzamento é baseado em árvores que representam os indivíduos. O cruzamento é descrito graficamente pelas Figuras 2.11 e 2.12.

Mais especificamente, o cruzamento baseado em árvores segue os seguintes passos:

- Selecionam-se dois indivíduos como pais utilizando-se alguns dos métodos de seleção descritos na subseção 2.3.5. Veja um exemplo de dois pais na Figura 2.11.
- Seleciona-se aleatoriamente uma subárvore de cada pai. A Figura 2.11 mostra as subárvores destacadas em negrito.
- Trocam-se as subárvores selecionadas entre os pais, obtendo-se como resultado os indivíduos filhos. Eles são exibidos pela Figura 2.12.

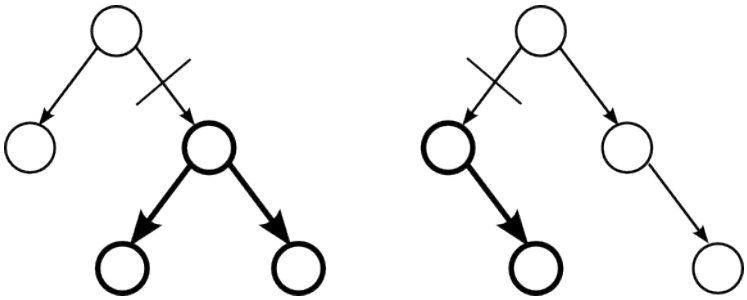

Figura 2.11: Dois indivíduos antes do cruzamento.

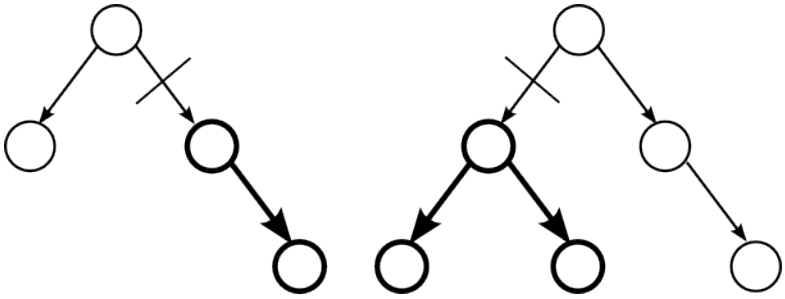

Figura 2.12: Dois filhos produzidos pelo cruzamento.

É comum utilizar probabilidade de cruzamento entre 50% e 95%. Quanto maior a probabilidade de cruzamento maior é a probabilidade de novas estruturas serem introduzidas na população, porém maior a probabilidade de boas estruturas serão tiradas da população. Se a probabilidade for baixa, a convergência do algoritmo poderá ser muito lenta.

Mutação

O operador de mutação é utilizado em somente um indivíduo. Normalmente depois do cruzamento, cada filho produzido sofre mutação com baixa probabilidade. A probabilidade de mutação é um parâmetro de execução do algoritmo e geralmente é ajustado de maneira empírica. A utilização desse operador tem por objetivo aumentar o espaço de busca, e evitar problemas como a falta de diversividade entre os indivíduos da população, introduzindo novas estruturas não introduzidas pela geração da população inicial.

Existem vários métodos para fazer mutação em PG, tais como: *point mutation, permutation, expansion mutation* e *subtree mutation*. O *point mutation* substitui um único nó da árvore por outro do mesmo tipo (função ou terminal) selecionado de forma aleatória. O método *permutation* troca as posições de dois nós de um mesmo indivíduo. O método *expansion mutation* troca um terminal da árvore por uma subárvore gerada de maneira aleatória. Já o método *subtree mutation* substitui uma subárvore do indivíduo por outra subárvore gerada de maneira aleatória (Banzhaf *et al.*, 1998; Freitas, 2002b). A Figura 2.13 exemplifica o processo de mutação de um ponto (*point mutation*) em um indivíduo de PG.

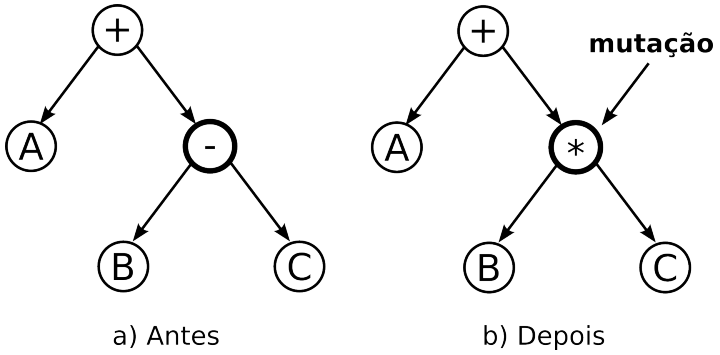

Figura 2.13: Ação do operador de mutação *point mutation* em um indivíduo.

Reprodução

A reprodução é um operador simples. Um indivíduo é selecionado e copiado, e a copia é colocada na população da próxima geração ficando duas versões do mesmo indivíduo na população (Banzhaf *et al.*, 1998).

2.3.7 Critério de Término

O critério de término é responsável por interromper o laço de repetição do processo evolutivo que, idealmente, não teria fim. Existem várias maneiras de interromper a evolução. Os mais comuns serão apresentados a seguir:

- **Número de gerações:** quando o algoritmo atingir um número máximo de gerações, especificado como parâmetro do algoritmo, o mesmo é interrompido.
- **Convergência:** quando não ocorre melhora de soluções por um longo número de iterações.
- **Valor da Função Objetivo** (para problemas de otimização): quando o valor da função objetivo é conhecido, o algoritmo é finalizado quando se encontra esse valor.

2.3.8 Um algoritmo básico de PG

Depois de definido os elementos de um indivíduo em PG (funções, terminais, métodos de seleção, operadores genéticos e inicialização da população) é possível executar o algoritmo. Há duas maneiras de executar o algoritmo, uma é utilizando a abordagem geracional e a outra é utilizando a abordagem *steady-state*. Na abordagem geracional, uma nova população é gerada a partir da população atual em um ciclo. A nova população substitui a atual ou parte dela (no caso do elitismo) e o ciclo continua. Na abordagem *steady-state*, não há gerações, apenas aplicação dos operadores na mesma população.

Antes de apresentar as duas abordagens é necessário entender os passos preparatórios para execução do algoritmo.

Passos preparatórios

Segundo Banzhaf *et al.* (1998) os passos preliminares para execução de um algoritmo de PG são:

- Definir o conjunto de terminais;
- Definir o conjunto de funções;
- Definir a função de avaliação (para calcular a aptidão);
- Definir parâmetros como tamanho da população, tamanho máximo do indivíduo, probabilidade de cruzamento, mutação, método de seleção e critério de término.

Uma vez que esses passos preliminares tenham sido definidos o algoritmo de PG pode começar a ser executado utilizando o método geracional ou *steady state*.

Método Geracional

Tradicionalmente PG utiliza o método geracional que em geral possui as gerações bem definidas e distintas. Cada geração é representada por uma população completa de indivíduos. A nova população é criada a partir da população atual. De acordo com Banzhaf *et al.* (1998), o ciclo de execução do método geracional segue os seguintes passos:

1. Inicialize a população;
2. Avalie os indivíduos pertencentes a população e atribua o valor de aptidão para cada um deles;
3. Até a nova população estar completamente formada, repita os seguintes passos:
   - Selecione os indivíduos da população utilizando o algoritmo de seleção;
   - Aplique os operadores genéticos sobre os indivíduos selecionados;
   - Insira os indivíduos resultantes dos operadores genéticos dentro da nova população.
4. Se o critério de término não foi alcançado, então continue. Caso contrário, substitua a população existente pela nova população e repita os passos de 2 a 4;
5. Apresente o melhor indivíduo da população como saída do algoritmo.

Método Steady-State

O método *steady-state* é a principal alternativa ao método geracional em PG. Essa abordagem não fixa intervalo de gerações, em vez disso há um fluxo contínuo de indivíduos reunidos no qual se combinam para gerar filhos. Os filhos substituem os indivíduos da mesma população. Um exemplo básico do método *setady-state* é apresentado a seguir (Banzhaf *et al.*, 1998):

1. Inicialize a população;
2. Escolha aleatoriamente um subconjunto da população para fazer parte do torneio (competidores);
3. Avalie a aptidão de cada competidor do torneio;

4. Selecione os vencedores do torneio utilizando o algoritmo de seleção;
5. Aplique os operadores genéticos sobre os vencedores do torneio;
6. Substitua os perdedores do torneio pelos filhos gerados a partir dos operadores genéticos;
7. Repita os passos de 2 a 7 até que o critério de término seja alcançado;
8. Apresente o melhor indivíduo da população como saída do algoritmo.

Essa abordagem é chamada *steady state* porque os operadores genéticos são aplicados de maneira assíncrona e não há mecanismo centralizado de gerações.

2.4 Algoritmos Culturais

Um Algoritmo Cultural (AC) é um algoritmo evolucionário baseado no processo de evolução cultural da humanidade (Reynolds, 1994). Os ACs foram propostos por Robert Reynolds (Reynolds e Detroid, 2002; Reynolds, 1999; Reynolds *et al.*, 2001; Reynolds e Saleem, 2001) como um complemento à metáfora evolutiva, utilizada na Computação Evolutiva, que se concentra nos aspectos genéticos da evolução e na teoria da seleção natural proposta por Darwin. Em contrapartida, os Algoritmos Culturais baseiam-se em teorias sociais e arqueológicas que modelam a evolução cultural dos povos (Becerra, 2002).

2.4.1 Inspiração Natural

Acredita-se que a cultura evolui ao longo das gerações. E, além disso, a evolução cultural acontece mais rapidamente que a evolução genética. Isso permite uma melhor adaptação ao ambiente do que a possível adaptação pela genética (Reynolds, 1992). Sugere-se que, ao longo dos tempos, o ser humano evoluiu um conjunto único de capacidades que permitiu a formação, codificação e transmissão de informações culturais. O conhecimento é codificado de forma a ser acessível a todos os indivíduos de uma sociedade. Uma vez codificado, esse conhecimento é assimilado por cada indivíduo da sociedade sob o prisma de suas experiências anteriores, podendo ou não esse indivíduo incorporar novos conhecimentos àqueles possuídos por sua sociedade.

As pesquisas a respeito da evolução cultural focam-se em dois níveis: microevolucionário e macroevolucionário. No nível microevolucionário se tem a modelagem da população em si. Já no nível macroevolucionário é modelado o conhecimento adquirido

pelos indivíduos ao longo das gerações e que, codificado e armazenado, ajuda a guiar o comportamento dos indivíduos em suas populações. É importante salientar que as informações culturais podem ser transmitidas tanto entre indivíduos de uma população quanto entre outras populações. Os algoritmos culturais têm por objetivo acelerar a taxa de convergência de Algoritmos Evolutivos, ou melhorar as populações geradas por meio desse mecanismo dual de herança.

Na próxima seção é apresentado o funcionamento geral de um Algoritmo Cultural.

2.4.2 Funcionamento Básico de um AC

Há um *framework*[3] básico proposto por Reynolds (1994) no qual são descritos dois componentes principais dos Algoritmos Culturais. Cabe salientar que não é preciso implementar todas as propriedades dos ACs, mas elas são apresentadas para que seja mantida a completude neste trabalho. A seguir são descritos os dois componentes principais de um Algoritmo Cultural:

- **Espaço Populacional:** conjunto de soluções que pode ser modelado utilizando qualquer técnica de inteligência artificial que faça uso de uma população de indivíduos;

- **Espaço de Crenças (*Mapa* do grupo):** local onde tem-se o armazenamento e representação do conhecimento (experiências ou *mapas* individuais) adquirido ao longo do processo evolutivo. É a partir desse conhecimento armazenado que os indivíduos são guiados na direção das melhores regiões do espaço de busca.

Os indivíduos são descritos por um conjunto de características e comportamentos e por um *mapa* que generaliza os conhecimentos e experiências adquiridos por esse indivíduo. Essas características e comportamentos são modificados por operadores genéticos que podem ser influenciados socialmente.

Da mesma maneira, os conhecimentos e experiências são unidos e modificados para formar o espaço de crenças. Essas unificações de modificações também são feitas por meio de operadores especiais.

Os símbolos utilizados para representar o conhecimento no espaço de crenças podem ser modificados ao longo das gerações. Dessa forma, é possível remover ou adicionar novas características e esquecer ou adquirir conhecimento por meio das experiências adquiridas pela população. A cada nova geração de indivíduos, é feita uma avaliação e o conhecimento

[3]Conjunto de conceitos usados para resolver um problema de um domínio específico.

dos melhores indivíduos pode ou não fazer parte do espaço de crenças. A nova população a ser gerada é influenciada com conhecimento anteriormente armazenado por meio de operadores.

Tanto o Espaço de Crenças quanto os indivíduos de uma população podem ser influenciados pelo que se chama Protocolos de Intercomunicação. A comunicação entre os indivíduos de uma geração e o Espaço de Crenças é dada por um protocolo dito Função de Aceitação. Já a interação do Espaço de Crenças com a população de indivíduos é chamado de Função de Influência.

O *framework* básico dos Algoritmos Culturais é dado pelo pseudo-código da Figura 2.14 e pela Figura 2.15.

Algoritmo Cultural
Inicializa a População
Inicializa o Espaço de Crenças
Repita
    Avalie a População
    Ajuste o Espaço de Crenças por meio da Função de Aceitação
    Gere a próxima População a partir da atual considerando a
        Função de Influência
Até que a condição de término seja alcançada

Figura 2.14: Algoritmo Cultural (Reynolds *et al.*, 2003)

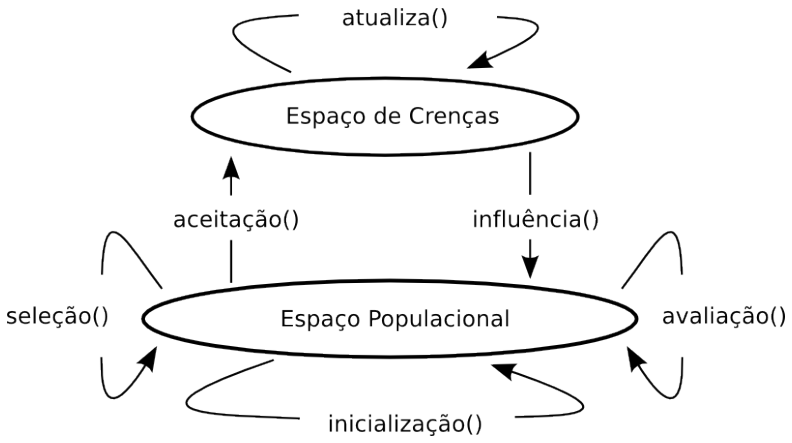

Figura 2.15: Funcionamento básico de um algoritmo cultural.

2.4.3 Principais Característica

De acordo com Reynolds *et al.* (2003), as principais características de um Algoritmo Cultural são:

- **Mecanismo dual de herança:** são herdadas características tanto em nível da população quanto em nível do espaço de crenças;
- **Evolução guiada por conhecimento:** tanto a população quanto o espaço de conhecimento podem ser organizados de forma hierárquica, permitindo a criação de nichos e, ao mesmo tempo, uma distribuição do conhecimento adquirido;
- **Conhecimento sobre o domínio separado dos indivíduos:** o conhecimento adquirido e armazenado no espaço de crenças é compartilhado entre os indivíduos. Assim, quando um indivíduo é eliminado da população, o conhecimento adquirido pelo mesmo permanece. Além disso, as técnicas de aquisição e manipulação de conhecimento podem ser adaptadas ao domínio da aplicação sem grandes mudanças na manipulação de indivíduos da população ou no algoritmo evolutivo utilizado. Essa característica permite a criação de ontologias para domínio sem conhecimentos prévios, ou seja, a criação *on-the-fly* (em tempo de evolução) da ontologia;
- **Suporte a auto adaptação em vários níveis:** permite tanto a auto adaptação da população quanto do conhecimento e da forma como o conhecimento é adquirido, ou seja, os parâmetros de controle, a representação, os operadores (tanto genéticos quanto sociais), a avaliação dos indivíduos e o protocolo de intercomunicação podem ser alterados a qualquer momento da evolução. Isso é particularmente útil para técnicas de meta evolução;
- **Diferentes taxas de evolução:** a evolução das populações e do conhecimento precisa ocorrer na mesma proporção. Segundo Reynolds (1992), o conhecimento é evoluído a uma taxa dez vezes maior que a população;
- ***Framework*:** permite a modelagem de diversas formas de evolução cultural, ou seja, a cultura pode mudar conforme as pesquisas em sociologia e biologia evoluírem a visão sobre como a aquisição e compartilhamento do conhecimento é adquirido pela sociedade (cultura).

O desenvolvimento de um Algoritmo Cultural pode ser dividido em três partes distintas: o desenvolvimento do componente de conhecimento, o desenvolvimento do componente populacional e os protocolos de intercomunicação.

2.4.4 Microevolução x Macroevolução

Os Algoritmos Culturais implementam mecanismo dual de herança. Esse mecanismo permite aos ACs explorarem tanto a microevolução quanto a macroevolução. A microevolução diz respeito à evolução que acontece em nível populacional. Já a macroevolução é a que ocorre sobre a cultura em si, ou seja, a evolução do espaço de crenças. Nos ACs a evolução ocorre de forma mais rápida do que nas populações sem o mecanismo de macroevolução.

Espaço Populacional

No Espaço Populacional são representadas as características e comportamentos dos indivíduos. Essa representação pode ser feita por meio de qualquer técnica que faça uso de uma população de indivíduos, como é o caso dos Algoritmos Genéticos e Programação Genética, entre outros.

Espaço de Crenças

O Espaço de Crenças é o repositório de símbolos que representa os conhecimentos adquiridos pelo Espaço Populacional ao longo do processo evolutivo. O Espaço de Crenças permite que os indivíduos sejam removidos da população sem que os conhecimentos por eles adquiridos sejam perdidos. O Espaço de Crenças foi criado para guiar os indivíduos em busca de melhores regiões.

Protocolos de Comunicação

Os protocolos de Comunicação ditam as regras sobre como os indivíduos que podem contribuir com conhecimentos para o Espaço de Crenças (Funções de Aceitação) e como o Espaço de Crenças influenciará novos indivíduos (Função de Influência), conforme segue:

- **Função de Aceitação:** Na Função de Aceitação são selecionados indivíduos que irão influenciar o Espaço de Crenças atual. Ela pode ser de dois tipos: estática ou dinâmica. Na estática pode utilizar-se do *ranking* absoluto (uma porcentagem da população é selecionada) ou do *ranking* relativo (os indivíduos com aptidão acima da média são selecionados). Já na função dinâmica, o percentual dos indivíduos selecionados varia ao longo do processo evolutivo. Inicialmente o processo é menos seletivo e torna-se mais restrito ao longo da evolução. Para Reynolds e Zhu (2001) a função de aceitação é a que mais influencia o desempenho de um Algoritmo Cultural.

- **Função de Influência:** A Função de Influência estabelece como o conhecimento armazenado no Espaço de Crenças interfere nos operadores do Espaço Populacional. Geralmente é utilizada a Função de Influência para cada tipo de conhecimento armazenado. A Função de Influência pode ser vista como um mecanismo de auto adaptação do processo evolutivo, já que ela adapta os operadores de acordo com o conhecimento adquirido. Normalmente é a Função de Influência que determina a direção e o tamanho das modificações impostas aos novos indivíduos.

2.4.5 Tipos de Conhecimento

Existem cinco maneiras de representar o conhecimento no Espaço de Crenças, as quais serão descritas a seguir.

Conhecimento Situacional

O conhecimento situacional contém um conjunto dos melhores indivíduos da população $< e_1, ..., e_n >$ em que n é o número de indivíduos pertencentes ao conhecimento situacional. Esse conhecimento representa um conjunto de exemplos para outros indivíduos da população. A estrutura de dados desse conhecimento é representada como uma lista de indivíduos, em que cada indivíduo dessa lista contém um valor para cada parâmetro e o valor da aptidão como na Figura 2.16.

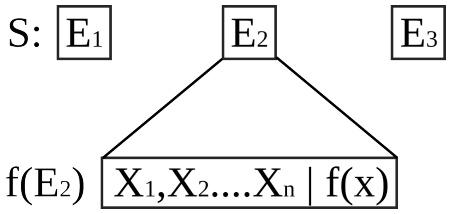

Figura 2.16: Representação do Conhecimento Situacional (Reynolds e Saleem, 2001).

Na Figura 2.16 E_i é um elemento da lista, $< X_1, X_2...X_n >$ são os valores de cada parâmetro do indivíduo e $f(x)$ é o valor da aptidão do indivíduo.

O conhecimento situacional é atualizado sempre que é encontrado um indivíduo cuja aptidão supere a aptidão do pior indivíduo pertencente ao conhecimento.

Conhecimento Normativo

O conhecimento normativo é representado por um conjunto de intervalos que caracterizam os intervalos que compõem as melhores soluções. Esses intervalos fornecem uma diretriz

para ajustes que podem ser feitos nos indivíduos. A estrutura de dados desse conhecimento é apresentado pela Figura 2.17.

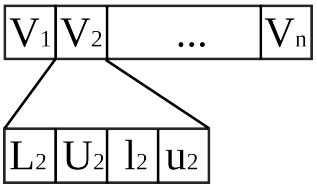

Figura 2.17: Representação do Conhecimento Normativo (Reynolds e Saleem, 2001).

Para cada variável V_i a estrutura de dados contém um limite superior e inferior l_i, u_i, e a aptidão de cada indivíduo que deu origem a esses limites, L_i e U_i respectivamente.

A atualização do conhecimento varia de acordo com o melhor indivíduo. Ou seja, se o indivíduo passou pela função de aceitação e seu limite inferior é maior ou o limite superior é menor que os limites armazenados no espaço de crença, então o intervalo é reajustado.

Conhecimento de Domínio

De acordo com Reynolds e Saleem (2001) esse conhecimento consiste da faixa de domínio para todos os parâmetros e os melhores indivíduos da população, similar a representação do conhecimento situacional. Entretanto, o propósito desse conhecimento é diferente. Para um dado problema é possível que se conheça alguma coisa a respeito da forma ou topologia do espaço de busca pertencente ao domínio do problema. Por exemplo, assuma-se que o espaço de busca do problema seja composto de muitos picos ou cones. Assim, o conhecimento sobre essas estruturas cônicas pode, por exemplo, ser utilizada para fazer previsões sobre qual direção e magnitude de ajustes e mudanças, buscando evitar ou não se limitar a esses picos.

Conhecimento Histórico

Esse conhecimento armazena importantes eventos durante o processo de busca. Foi originalmente proposto para funções objetivo dinâmicas no qual foi utilizado para encontrar padrões nas mudanças ambientais. Indivíduos guiados pelo conhecimento histórico podem consultar aqueles eventos armazenados para guiar suas decisões sobre qual direção seguir.

A estrutura utilizada para representar o conhecimento histórico pode ser vista pela Figura 2.18, em que e_i representa o melhor indivíduo encontrado antes da i-ésima alteração do ambiente. ds_i é a distância média das mudanças para a característica i e dr_i é a direção média se existem mudanças para a característica i.

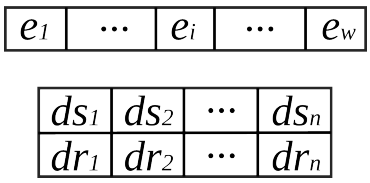

Figura 2.18: Representação do Conhecimento Histórico (Becerra e Coello, 2005).

Conhecimento Topográfico

O objetivo do conhecimento topográfico é dividir o espaço de busca em células com um certo nível de granularidade de acordo com características especiais buscando guiar os indivíduos da população em direção as melhores células. Cada célula mantém uma faixa de indivíduos e o seu estado pode ser medido pela média das aptidões dos indivíduos pertencentes a ela. As células podem ser recursivamente subdivididas em células menores buscando refinar o espaço de busca, ou se manter inalterado durante todo o processo evolutivo (Jin e Reynolds, 1999; Xue e Guo, 2007).

Para influenciar a geração de novas soluções, o conhecimento topográfico seleciona as células que possuam as boas soluções e busca produzir indivíduos próximos a essas células/regiões. Para atualização desse conhecimento pode-se dividir a célula em k células menores sempre que se encontra um novo indivíduo que seja melhor que o melhor indivíduo da célula (Becerra e Coello, 2005) ou atualizar a célula com o indivíduo caso seja melhor que o pior indivíduo da célula.

2.5 Mineração de dados em Planos de Saúde

No Brasil com a criação da ANS (Agência Nacional de Saúde Suplementar) em novembro de 1999 pela medida provisória n. 1926, convertida na lei n. 9.961, como órgão de regulação, normatização, controle e fiscalização das atividades que garantem a assistência suplementar à saúde ocorreram mudanças na estrutura do setor de saúde suplementar, no qual agora o modelo de atenção é focado nas ações de promoção e prevenção de doenças e não mais centrado na doença, além de prever uma ampliação na utilização de sistemas de informação como insumo estratégico (Dallagassa *et al.*, 2009).

Com a utilização desses sistemas de informação o acúmulo de dados é cada vez maior. Sendo a otimização dos processos operacionais cada vez mais um fator crítico para o sucesso das organizações que atuam na área de saúde, a tecnologia de informação ocupa vez mais um papel de destaque entre as ferramentas de gerenciamento na gestão da saúde.

Uma dessas ferramentas de auxílio ao gerenciamento na gestão da saúde é a mineração de dados.

Existe um grande potencial na aplicação de mineração de dados em planos de saúde. Geralmente essas aplicações podem ser agrupadas como (I) avaliação da eficácia do tratamento, (II) gestão em saúde, (III) gestão de relacionamento com o beneficiário e (IV) detecção de fraudes e abuso (Koh e Tan, 2005; Obenshain, 2004), como detalhado a seguir.

2.5.1 Avaliação da eficácia do tratamento

Aplicações de MD podem ser desenvolvidas para avaliar a eficácia de tratamentos médicos. Comparando e contrastando causas, sintomas e andamento dos tratamentos, a MD pode fornecer uma análise de quais direções e ações se mostram mais eficazes (Milley, 2000 apud Koh e Tan, 2005). Por exemplo, os resultados de grupos de pacientes tratados com diferentes combinações de medicamentos para a mesma doença ou condição, podem ser comparados para determinar qual tratamento possui o melhor custo-benefício.

Pode ser desenvolvido também ferramentas para ajudar a identificar padrões de tratamentos para doenças específicas. Uma aplicação para tratamentos inclui a associação de vários efeitos colaterais de um tratamento, agrupando sintomas comuns para apoiar o diagnóstico, determinando quais os melhores remédios (drogas) para o tratamento de uma determinada sub-população que responde de maneira diferente da população em geral para certos remédios, e estabelecer assim os tratamentos que podem reduzir o risco de sofrimento (Milley, 2000 apud Koh e Tan, 2005).

2.5.2 Gestão em Saúde

Com o objetivo de apoiar a gestão em saúde, aplicações de mineração de dados podem ser desenvolvidas para melhor identificar e controlar estados de doenças crônicas e pacientes de alto risco, designar intervenções apropriadas e reduzir o número de admissões em hospitais (Koh e Tan, 2005).

Por exemplo, a *American Healthways* (Obenshain, 2004) presta serviços de gestão em doenças, como diabetes para hospitais e planos de saúde, destinados a melhorar a qualidade e reduzir custos do tratamento de indivíduos com essa doença. Para aumentar a capacidade da empresa em identificar pacientes de alto risco, a *American Healthways* utiliza a tecnologia de modelagem preditiva (classificação e regressão). Uma grande quantidade de informações sobre pacientes é combinada e explorada para prever a probabilidade de ocorrer problemas de curto prazo em relação a saúde do paciente, e

assim intervir de maneira proativa para obter melhores resultados de curto e longo prazo (Obenshain, 2004).

Um modelo robusto construído pelo processo de mineração de dados pode identificar pacientes que possuem tendência a condição de alto risco. Esse tipo de informação dá aos coordenadores de cuidados em enfermagem um avanço na identificação de pacientes de alto risco facilitando a tomada de decisão para melhorar a qualidade do atendimento dos pacientes e prevenir problemas de saúde no futuro. Por exemplo, Cavalcante (2009) desenvolveu um estudo que teve como objetivo identificar e avaliar a prevalência dos fatores de risco cardiovasculares e a associação entre eles, presentes em pacientes com obstrução das artérias coronárias.

A MD também pode melhorar os cuidados com saúde e diminuir custos evitando complicações clínicas, desenvolvendo melhores práticas de atendimento, tratamento e diagnóstico, diminuindo o tempo de permanência dos paciente no hospital e fornecendo mais informações aos médicos. Guimarães *et al.* (2006) realizou um estudo para melhorar o processo de diagnóstico da Síndrome da Apneia Obstrutiva do Sono, que quando não diagnosticado reflete em custos elevados para as instituições de saúde, além de possuir um processo de diagnóstico difícil e caro.

2.5.3 Gestão de Relacionamento com o Beneficiário

A aplicação de gestão de relacionamento com o beneficiário foca em determinar e satisfazer as necessidades do beneficiário. Ela envolve o rastreamento de informações dos beneficiários, tais como dieta, dados sobre exercícios, informações sobre diagnósticos anteriores, histórico da família e informações sobre alergias. Com essas informações, os planos de saúde podem enviar *e-mails* aos beneficiários sobre estudos de saúde recém-publicados que podem ser de interesse do beneficiário ou oferecer prescrições especializadas que podem se ajustar ao perfil do beneficiário (Siau, 2003).

Para obter maior sucesso no contexto do relacionamento com o beneficiário pode-se desenvolver a MD para determinar as preferências do beneficiário, encontrar padrões de tratamento e necessidades atuais e futuras. Essas características são utilizadas para melhorar o nível de satisfação dos beneficiários. As interações com os beneficiários podem ocorrer por meio de *call centers*, consultórios médicos, departamentos de faturamento, em internamentos e ambulatórios (Koh e Tan, 2005).

Pode-se também utilizar a MD para identificar beneficiários propensos a utilizar serviços de saúde específicos, ou encontrar os beneficiários que mais podem beneficiar-se de serviços específicos, como um tratamento preventivo que pode produzir uma significativa

redução na utilização futura dos serviços de saúde. Encontrar os beneficiários que mais precisam de um cuidado específico deve ser feito constantemente, assim é possível melhorar os canais de comunicação com os beneficiários apropriados, objetivando melhorar a saúde, relacionamento de longo prazo e a lealdade do beneficiário.

Melhorar os serviços prestados também é uma maneira de relacionamento com o beneficiário e a MD pode revelar possíveis maneiras de melhorar o serviço. Por exemplo, Vianna e Moro (2005) estudam as características relacionadas a mortalidade infantil, buscando reconhecer o desempenho dos serviços de saúde prestados, focando-se principalmente nas mulheres e crianças. Com os conhecimentos encontrados no processo de MD, Vianna e Moro (2005) objetivam melhorar os serviços de saúde prestados.

No contexto do relacionamento com o beneficiário pode-se utilizar a MD para *marketing* em saúde. Rafalski (2002) utiliza técnicas de MD como oportunidade para melhorar a continuidade dos tratamentos, melhorando assim a satisfação do beneficiário e melhorando as receitas do *Sinai Health System* de Chicago. Nesse trabalho Rafalski (2002) analisa a conformidade das pacientes no pré-natal entre as clínicas afiliadas utilizadas para os cuidados iniciais e o hospital utilizado para o parto. Isso levou ao desenvolvimento de uma pesquisa telefônica utilizada para determinar por que as pacientes que estavam recebendo cuidados pré-natais em um sistema de saúde escolheram entregar seus bebês a outros hospitais não filiados. Os resultados foram utilizados pela administração e pelo *marketing* a fim de melhorar os processos e minimizar a perda de negócios.

2.5.4 Detecção de Fraudes e Abuso

Aplicações de MD que procuram detectar fraudes e abusos frequentemente estabelecem normas e então identificam padrões incomuns ou anormais sobre solicitações por parte dos médicos, laboratórios, clínicas e outros. Entre outras coisas, essas aplicações podem destacar prescrições ou encaminhamentos inapropriados, seguros e solicitações médicas fraudulentos.

Por exemplo, Ortega *et al.* (2006) descreve um sistema para detecção de fraudes e abusos de reivindicações médicas baseado em mineração de dados, utilizado por uma empresa de seguro saúde do Chile. O sistema conseguiu detectar fraudes a uma taxa de aproximadamente 75 casos fraudulentos e abusivos por mês, realizando as detecções 6,6 meses antes que um humano. Durante o ano de 2005 a empresa de seguro saúde foi capaz de estimar que o novo sistema de detecção de fraude rejeitou reivindicações de saúde que contribuíram entre 9,5% e 10% do total bruto dos custos.

2.6 Considerações Finais

Este capítulo apresentou a revisão bibliográfica necessária para o entendimento desta dissertação. Com o acumulo significativo de informações, a utilização de algoritmos e ferramentas de MD torna a descoberta de conhecimento oculto em base de dados possível, dando suporte à tomada de decisões.

Para encontrar esses conhecimentos busca-se resolver a tarefa de classificação em MD e para isso escolheu-se trabalhar com algoritmos evolucionários, especificamente com programação genética que é uma técnica de busca e otimização e tem sua inspiração em conceitos da teoria de seleção natural das espécies proposta por Darwin. Além disso aplica-se à essa técnica o algoritmo cultural como um complemento a metáfora evolutiva.

Espera-se com essa revisão bibliográfica dar suporte ao entendimento da investigação da viabilidade da aplicação de um algoritmo cultural (fundamentado na combinação de programação genética e algoritmos culturais), no contexto de planos de saúde suplementar, com o objetivo de descobrir conhecimento oculto e interessante (surpreendente, novo) em banco de dados da área da saúde utilizando impressões gerais (conhecimento subjetivo), fornecidas pelo usuário, como conhecimento de domínio no algoritmo cultural.

Capítulo

3

Programação Genética Aplicada à Mineração de Dados

3.1 Considerações Iniciais

Como apresentado no Capítulo 2 a PG é um poderoso método de busca baseado no paradigma de algoritmos evolucionários, que consiste em uma população de "programas" que são soluções candidatas para resolver um problema específico.

Neste capítulo apresenta-se a abordagem utilizada no desenvolvimento do algoritmo de PG usado neste trabalho, com o objetivo de encontrar padrões em dados de procedimentos realizados por beneficiários de planos de saúde suplementar, buscando a previsão da propensão para determinados tipos de doenças em beneficiários com características semelhantes aos padrões encontrados, além de apresentar uma nova abordagem para "boolenização" de atributos contínuos no contexto de PG. Para isso utilizou-se dados pre-processados, conforme descrito no Capítulo 5.

3.2 Modelo do Sistema

Como apresentado no Capítulo 2, PG é um método que consiste em uma população de "programas" que representam soluções candidatas para resolver um problema específico. Os programas, denominados indivíduos, são representados na forma de árvores, em que os nós internos são funções e os nós folhas são variáveis ou constantes (Koza, 1992). As

funções e os nós folhas são separados em dois conjuntos. Nas seções seguintes são descritos: a maneira como os conjuntos foram estabelecidos, a representação dos indivíduos, o funcionamento dos operadores genéticos e os demais mecanismos do sistema.

3.2.1 Funções e Terminais

Segundo Koza (1992), um indivíduo (solução candidata) em PG pode conter não somente valores de atributos, mas também funções. Portanto, para criar novos indivíduos em PG é necessário determinar dois conjuntos: o conjunto de funções (Koza, 1992) ou não-terminais (Montana, 1995) e o conjunto de terminais.

O problema do fechamento para descoberta de regras e terminais

Os dados utilizados nesta dissertação estão organizados no formato de uma tabela, nos quais existem atributos cujo domínio pode ser de valores categóricos (nominal) ou contínuos (real). Por padrão, a PG não foi desenvolvida para tratar diferentes tipos de dados ao mesmo tempo, assim, a propriedade de fechamento não é satisfeita. Diferentes atributos são associados a diferentes funções ocasionando problemas. Por exemplo: $Idade < 18$ é válido, mas $Sexo < Feminino$ não é.

Segundo Freitas (2002b), predição de regras são essencialmente combinações lógicas de atributo-valor. Assim a propriedade de fechamento pode ser satisfeita "boolenizando" todos os terminais do conjunto de terminais e usando somente funções booleanas no conjunto de funções. Dessa forma todos os nós da árvore possuirão como saída um valor booleano (*verdadeiro* ou *falso*) que pode ser utilizado como entrada por qualquer outro nó da árvore.

Foi necessário efetuar o processo de "boolenização" para os atributos categóricos e contínuos. Para "boolenização" de atributos categóricos utilizou-se a abordagem de particionamento proposto por Freitas (2002b).

Para fazer o particionamento em um atributo categórico $Atrib_i$ que possui como domínio V_i valores distintos é feita uma partição desse atributo em V_i novos atributos booleanos, cada um deles definido pela condição "$Atrib_i = Val_{ij}$", em que Val_{ij} é o j-ésimo valor do domínio do atributo $Atrib_i$. Como exemplo tem-se o atributo $Estado$, que após a "boolenização" foi possível criar terminais como $Estado = SC, Estado = PR, ..., Estado = GO$.

Para atributos contínuos, Freitas (2002b) sugere uma abordagem no qual o conjunto de valores do atributo é particionado em k ($k > 2$) subconjuntos mutuamente exclusivos e para cada um desses subconjuntos é criado um novo atributo booleano. Essa abordagem

é pouco utilizada, principalmente por ser mais complexo e mais caro computacionalmente que a partição binária ($k = 2$). Outra abordagem é utilizar ganho de informação para fazer o particionamento. Mais detalhes dessa abordagem podem ser encontradoss em (Hu, 1998). Nesta dissertação propõe-se uma abordagem para "boolenização" de atributos contínuos utilizando lógica difusa (Zadeh, 1965), que será apresentada em mais detalhes a seguir.

Lógica difusa pode representar informação de acordo com o grau de verdade e permite codificar expressões do tipo *"mais ou menos"*, *"aproximadamente"*, *"quase"*, *"pouco"*, *"muito"*, etc. A utilização da lógica difusa na "boolenização" de atributos contínuos elimina a necessidade de alterações desses atributos na fase de pré-processamento. Na abordagem proposta, a lógica difusa foi empregada de maneira relativamente simples, apenas para validação das regras dentro da PG juntamente com os atributos categóricos.

Apesar de adotar a abordagem simples da lógica difusa, a eficiência da máquina de inferência difusa depende da forma das funções de pertinência. As formas mais empregadas são: triangular, trapezoidal e gaussiana. Em muitos projetos a forma escolhida é fixa durante toda a vida útil do sistema. Nesta dissertação adotou-se para cada atributo contínuo três termos linguísticos, *"baixo"*, *"médio"* e *"alto"*. Esses termos são representados por três funções de pertinência no formato trapezoidal (Romão, 2002), conforme pode ser visto na Figura 3.1.

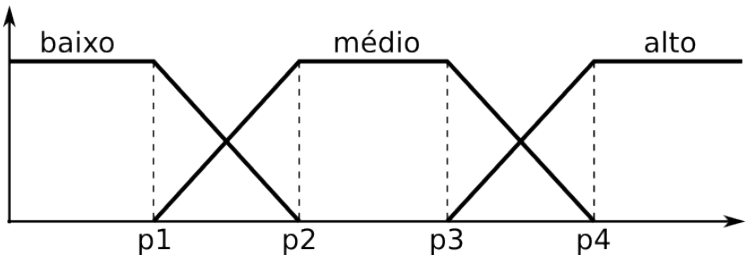

Figura 3.1: Conjuntos difusos para três valores linguísticos

Tomando a base dos trapézios como referência, e impondo-se algumas restrições, é possível representar os três valores linguísticos de cada atributo com apenas quatro parâmetros: $p1$, $p2$, $p3$ e $p4$. A simetria entre os trapézios é dada pelas seguintes coordenadas:

Ponto de interseção das curvas baixo e médio:

$$\frac{p1 + p2}{2}$$

Ponto de interseção das curvas médio e alto:

$$\frac{p3 + p4}{2}$$

Para definição dos valores de $p1$ à $p4$, adotou-se uma abordagem que permite a adaptação pelo usuário do tamanho da função de pertinência, ou seja, o formato será sempre trapezoidal mas o usuário poderá escolher a extensão de cada conjunto difuso, incorporando assim conhecimento do usuário sobre o domínio da aplicação.

Sendo assim, sempre que um atributo $Atrib_i$ for contínuo (numérico) terá terminais do tipo "$Atrib_i = baixo$", "$Atrib_i = médio$" e "$Atrib_i = alto$", em que os termos dos parâmetros difusos são definidos pelo usuário. Essa abordagem torna o processo de "boolenização" dos atributos contínuos mais simples e incorpora conhecimento do usuário sobre o domínio da aplicação. Como exemplo tem-se o atributo $Consulta$ que se refere ao número de consultas realizadas por um determinado beneficiário, que gerou os atributos booleanos $Consulta = baixo$, $Consulta = médio$ e $Consulta = alto$.

Dessa maneira foi possível determinar o conjunto de terminais (T) e de funções (F):

$$T = (Sexo = Masculino, ..., Idade = 10_A_19, ..., Consulta = Baixa, ...) \tag{3.1}$$

$$F = (E, OU) \tag{3.2}$$

3.2.2 Representação do Indivíduo

Para definir a representação de um indivíduo em PG na tarefa de classificação é necessário escolher entre a abordagem *Michigan* ou *Pittsburg*. Na abordagem *Michigan* um indivíduo corresponde a uma única regra de classificação de modo que um conjunto de indivíduos é uma solução completa para o problema de classificação. Já na abordagem *Pittsburg* um indivíduo corresponde a um conjunto de regras representando assim uma solução completa para o problema de classificação (Freitas, 2002a,b).

Nesta dissertação utilizou-se uma abordagem híbrida, na qual cada indivíduo pode possuir várias regras de classificação, mas esse indivíduo não representa uma solução completa para o problema de classificação, pois todos os indivíduos da população possuem o mesmo consequente, ou seja, preveem a mesma classe. Portanto, para gerar uma solução completa para um problema de classificação é necessário executar o algoritmo para cada classe. A seguir apresenta-se a formalização da representação do indivíduo.

Com a definição do conjunto de terminais e funções é possível criar árvores como a da Figura 3.2.

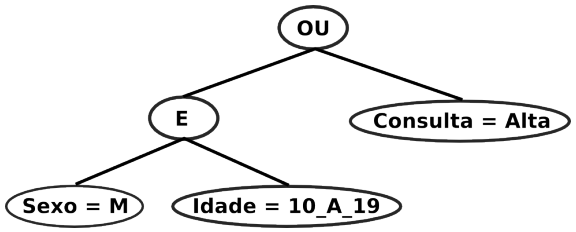

Figura 3.2: Representação do Indivíduo

A árvore da Figura 3.2 representa a parte dos antecedentes das regras (parte SE):

$$SE\ ((Sexo = M)\ E\ (Idade = 10_A_19))\ OU\ (Consulta = Alta) \tag{3.3}$$

O consequente da regra (parte do ENTÃO) é escolhido previamente para todos os indivíduos da população. Nesse caso, um indivíduo pode ser considerado como um conjunto de regras na Forma Normal Disjuntiva (FND), ou seja, uma disjunção de regras no qual cada regra é uma conjunção de condições. Como exemplo, considere como consequente o atributo-valor "$Internação = mais_que_uma$" do indivíduo da Figura 3.2. Utilizando a FND ele pode ser decomposto formando assim duas regras que podem ser vistas a seguir:

$$SE\ ((Sexo = M)\ E\ (Idade = 10_A_19))\ ENTÃO\ (Internacão = mais_que_uma) \tag{3.4}$$

$$SE\ (Consulta = Alta)\ ENTÃO\ (Internacão = mais_que_uma) \tag{3.5}$$

Para facilitar a conversão na FND de uma árvore em PG utilizou-se uma restrição sintática. A restrição utilizada proíbe que um nó OU seja filho de um nó E. Com essa restrição fica garantido que as árvores geradas poderão ser convertidas pela FND.

Outra restrição importante é a restrição que busca melhorar a compreensibilidade das regras encontradas, em que não é permitido regras no indivíduo com mais de 5 antecedentes. Essa restrição é adotada na inicialização da população e nos operadores genéticos, assim como nas demais restrições.

Restrição Semântica

Segundo Freitas (2002b) a "boolenização" de todos os terminais satisfaz o requerimento de fechamento e garante a geração de árvores sintaticamente válidas. Entretanto, ainda é necessário tomar cuidado para evitar a geração de árvores semanticamente inválidas, tal como a árvore da Figura 3.3. Para evitar esse tipo de problema foi incorporado uma restrição na geração da população inicial e nos operadores genéticos para não permitir a geração de indivíduos com essa característica.

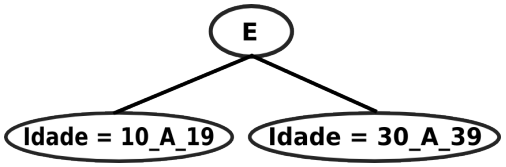

Figura 3.3: Indivíduo Semanticamente Inválido

Geração da população Inicial

Para geração da população inicial utilizou-se um algoritmo chamado *Probabilistic Tree-Creation* (PTC1) proposto por *Luke* Luke (2000). Esse algoritmo permite determinar parâmetros como a expectativa de tamanho da árvore gerada. Esse algoritmo não tenta gerar árvores completamente uniformes, em vez disso garante ajustes de parâmetros que outros algoritmos de inicialização, como o *GROW* (Koza, 1992), não permitem (e.g.: determinar a probabilidade de uma função aparecer em uma árvore).

Os parâmetros do PTC1 são E_{tree}, que indica o tamanho esperado para a árvore, D, que indica a profundidade máxima da árvore, q_n, que indica a probabilidade de escolha de uma função para quando o algoritmo precisar de uma função e q_t que indica a probabilidade de escolha de um terminal para quando o algoritmo precisar de um terminal.

A profundidade máxima foi definida como 4 e o tamanho esperado como 31. As probabilidades q_n e q_t foram distribuídas igualmente para terminais e funções. Entretanto, em experimentos foi possível verificar problemas relativos aos atributos que possuem domínio muito grande. Por exemplo, o atributo município possui como domínio 273 cidades e como a probabilidade era distribuída de maneira igualitária sobre todo o conjunto de terminais havia uma prevalência do município em relação aos outros atributos.

Procurando resolver o problema de prevalência de atributo adotou-se uma nova maneira de distribuir as probabilidades para o conjunto de terminais. Agora em vez de distribuir as probabilidades igualmente sobre todos os terminais, as probabilidades são distribuídas por atributos e depois a probabilidade de cada atributo é dividida pelo

domínio de cada atributo. Assim, com 55 atributos selecionados a probabilidade de cada atributo será de 1/55 e para cada um desses atributos divide-se sua probabilidade pelo seu domínio. Por exemplo, para o atributo MUNICÍPIO cada terminal gerado terá probabilidade de $(1/55) * (1/273)$. Para o atributo INTERNAÇÃO, que possui como domínio 3 valores, terá para cada terminal gerado a probabilidade de $(1/55) * (1/3)$. Com essa abordagem todos os atributos possuem a mesma chance de ocorrerem nas árvores geradas.

3.2.3 Operadores Genéticos

A evolução dos indivíduos ocorre transformando a população inicial e as novas populações por meio dos operadores genéticos. Neste algoritmo foram utilizados três operadores genéticos: cruzamento, mutação e reprodução. A seguir demonstra-se o funcionamento desses operadores.

Operador de Cruzamento

O operador de cruzamento troca material genético entre dois indivíduos da população. Para selecionar esses dois indivíduos utilizou-se o método de torneio. Nesse método de seleção k indivíduos são escolhidos aleatoriamente da população corrente. Os indivíduos escolhidos são comparados entre si, caracterizando um torneio, e aquele com maior valor da função de aptidão é selecionado para reprodução (Back *et al.*, 1999).

Após selecionar os dois pais (indivíduos) o cruzamento é feito de acordo com uma probabilidade preestabelecida trocando uma subárvore, selecionada de forma aleatória do primeiro pai, por uma subárvore selecionada do segundo pai também de forma aleatória. As Figuras 3.4 e 3.5 exemplificam esse processo.

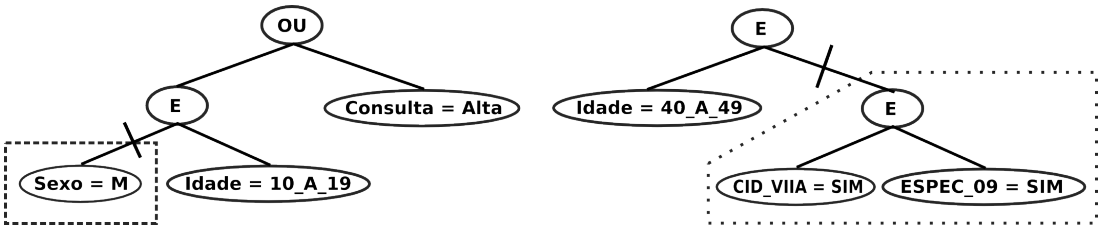

Figura 3.4: Seleção de subárvores dos pais

Após a geração dos novos indivíduos (filhos) a aptidão desses indivíduos é recalculada. Segundo Banzhaf *et al.* (1998) o operador de cruzamento em PG possui um efeito destrutivo nos indivíduos. Para resolver esse problema aplicou-se o método *brood recombination*(Banzhaf *et al.*, 1998) que é baseado no fato de que muitas espécies de

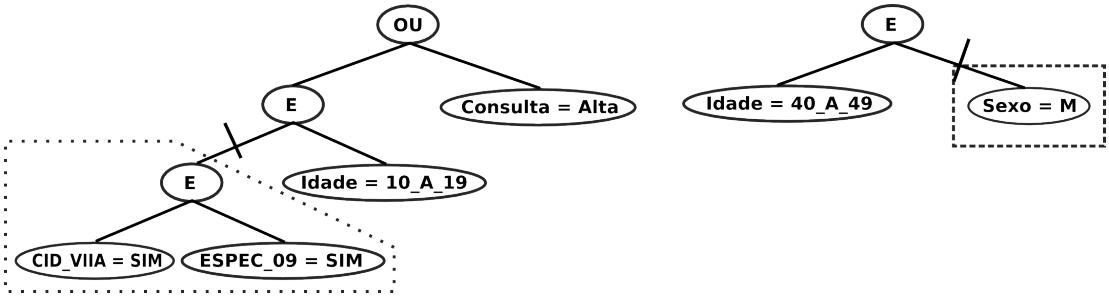

Figura 3.5: Troca de subárvores dos pais gerando novos indivíduos

animais produzem vários descendentes e que nem todos sobrevivem. Nesse método o excesso de descendentes é eliminado. O método aplicado neste algoritmo faz quatro cruzamentos produzindo oito descendentes dos quais são selecionados os dois indivíduos diferentes e de melhor aptidão.

Operador de Mutação

O operador de mutação é aplicado com uma probabilidade relativamente menor que a probabilidade de cruzamento e atua sobre um único indivíduo. Existem várias técnicas para se fazer mutação em PG tais como *point mutation*, *collapse mutation*, *expansion mutation*, *subtree mutation*, etc. (Banzhaf *et al.*, 1998).

A técnica escolhida foi a *point mutation*, a forma de mutação mais simples, que consiste em trocar um nó da árvore por outro randomicamente gerado e do mesmo tipo. "Mesmo tipo" significa que se o nó selecionado for um nó interno, ou seja, um nó que teve origem no conjunto de funções, então ele deve ser substituído por outro nó do mesmo conjunto de funções. A Figura 3.6 ilustra o processo de mutação de nó interno utilizado neste algoritmo.

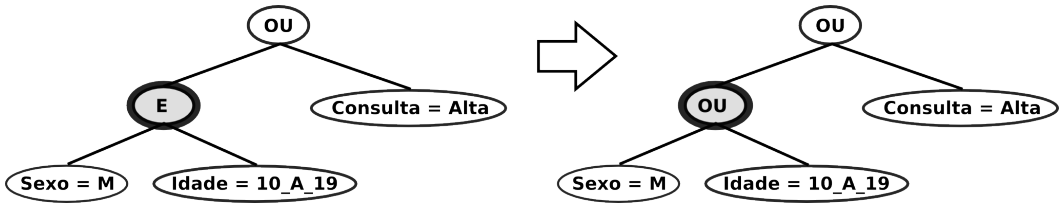

Figura 3.6: Mutação sobre um nó interno (função)

Se o nó selecionado for um nó folha, ou seja, um nó pertencente ao conjunto de terminais, então ele deve ser substituído por outro nó do mesmo conjunto de terminais. A Figura 3.7 exemplifica o processo de mutação de terminais utilizado neste algoritmo.

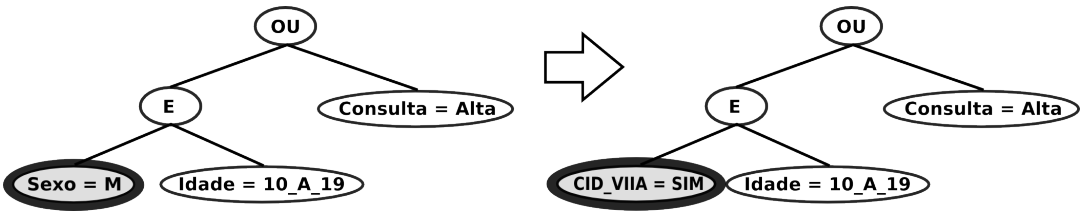

Figura 3.7: Mutação sobre um nó folha (terminal)

Operador de reprodução

O operador de reprodução é o operador mais simples entre os três aplicados neste algoritmo. Alguns indivíduos da população corrente são selecionados de maneira aleatória e inseridos na população da próxima geração.

3.2.4 Elitismo

A técnica de elitismo é utilizada para evitar a perda dos **melhores** indivíduos durante o processo de evolução. Permite que os n melhores indivíduos da população passem para a próxima geração sem sofrer alteração ou destruição. Esses indivíduos são selecionados para passar para a próxima geração, mas servem aos operadores genéticos, ou seja, eles também podem fornecer material genético para criação de novos indivíduos.

3.2.5 Avaliação das Regras (Função Fitness)

A função de aptidão (*fitness*) avalia a capacidade de adaptação de um indivíduo ao ambiente. Para descoberta de regras, a função de aptidão deve representar a qualidade da regra. Na tarefa de classificação em MD, para definir a qualidade da regra é necessário conhecer alguns conceitos básicos sobre avaliação de regras. Como exemplo, considere um caso simples, ou seja, um problema de classificação em que há somente duas classes para serem previstas, referenciadas como classes "positiva" e classes "negativa". Nesse caso, a matriz de confusão será uma matriz 2x2 como ilustrado pela Figura 3.8. Nessa figura, c denota a classe prevista pela regra, que é também chamada classe positiva. A classe prevista para uma dada instância é c se a instância satisfaz o antecedente da regra, denotado por A. Caso contrário a instância é atribuída a classe negativa (Bojarczuk *et al.*, 1999; Freitas, 2002b).

Os rótulos de cada quadrante da matriz possuem o seguinte significado:

- Verdadeiro positivo - número de instâncias que satisfazem A e tem a classe c.

- Falso positivo - número de instâncias que satisfazem A mas não tem a classe c.
- Verdadeiro negativo - número de instâncias que não satisfazem A e não tem a classe c.
- Falso negativo - número de instâncias que não satisfazem A mas tem a classe c.

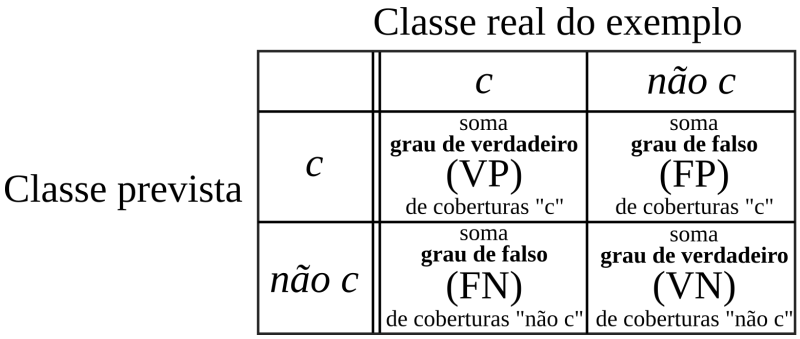

Figura 3.8: Matriz de confusão difusa

Denota-se vp, fp, vn e fn como o número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, respectivamente, no qual uma regra é utilizada para classificar um conjunto de casos. Nesta dissertação, para o critério de qualidade em regras, utilizou-se uma matriz de confusão para cada regra descoberta. Diferentemente dos métodos tradicionais, em que se tem C classes e a matriz de confusão teria tamanho CxC, aqui assume-se que há somente duas classes, produzindo uma matriz de confusão 2x2. Nesse caso, a classe c prevista pela regra é considerada como classe positiva, todas as outras classes são simplesmente consideradas como classes negativas e não existe distinção entre diferentes tipos de classes negativas.

Quando o atributo é difuso a soma considera o grau de verdadeiro (ou falso) de cobertura, conforme a matriz de confusão difusa da Figura 3.8 adaptada a partir da matriz de confusão da Figura 2.5.

Para ilustrar o funcionamento dessa matriz de confusão, considere um exemplo de como calcular uma matriz de confusão difusa para a regra:

$$(Consulta = Baixo), (SADT = Alto) \rightarrow (Internacão = mais_que_uma) \tag{3.6}$$

Supondo que o usuário forneceu, como significado do termo difuso “Baixo” do atributo Consulta, o conjunto difuso dado pela Figura 3.9.

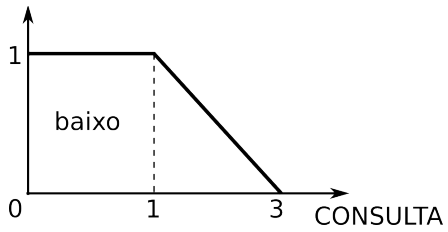

Figura 3.9: Função de pertinência (FP) para Consulta = "Baixo"

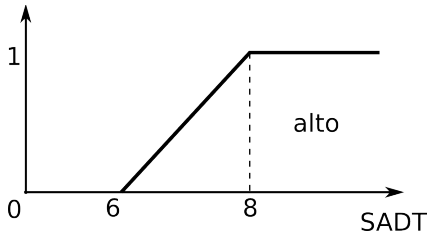

Figura 3.10: Função de pertinência para SADT = "Alto"

Supondo ainda que o usuário tenha definido o termo "Alto" do atributo SADT conforme o conjunto difuso da Figura 3.10.

Considerando também que a regra apresentada na Equação 3.6 possui implicitamente uma conjunção, é necessário adaptá-la para uma conjunção difusa representada pela interseção mínima, na qual empregam-se métodos de raciocínio aproximado (implicação) para derivar conclusões a partir da interseção dos fatos apurados na primeira etapa com os antecedentes das regras difusas, ou seja, $E\ difuso = min(cond1, cond2, ...)$ (Romão, 2002).

Tabela 3.1: Conjunto de dados fictício

Reg.	Consulta	$\mu_{Consulta}$	SADT	μ_{SADT}	min	Classe (Internação)
1	1	1.0	9	1.0	1.0	mais_que_uma
2	2	0.5	7	0.5	0.5	mais_que_uma
3	5	0.0	3	0.0	0.0	nenhuma
4	2	0.5	8	1.0	0.5	uma

Considere também o conjunto de dados da Tabela 3.1, incluindo-se o cálculo das FP's (Função de Pertinência), onde: μ_{CON} = FP do atributo Consulta; μ_{SADT} = FP do atributo SADT; mim = cálculo do E difuso, dado pela implicação mínima mostrada na Equação 3.7 (Mendel, 1995).

$$\mu_{A \to B}(x, y) = min[\mu_A(x), \mu_B(y)] \tag{3.7}$$

onde a função de pertinência $\mu_{A \to B}(x, y)$ mede o grau de verdade da relação de implicação entre x e y. O grau de pertinência $\mu_{CON}(x)$ e $\mu_{SADT}(x)$ são calculados conforme as Equações 3.8 e 3.9, baseados nas funções de pertinência das Figuras 3.9 e 3.10.

$$\mu_{CON}(x) = \begin{cases} 1, & se \quad x \leq 1 \\ 0, & se \quad x \geq 3 \\ \frac{3-x}{2}, & se \quad 1 < x < 3 \end{cases} \tag{3.8}$$

$$\mu_{SADT}(x) = \begin{cases} 1, & se \quad x \geq 8 \\ 0, & se \quad x \leq 6 \\ \frac{x-6}{2}, & se \quad 6 < x < 8 \end{cases} \tag{3.9}$$

Para calcular a matriz de confusão toma-se o primeiro registro do conjunto de dados: Consulta = 1, SADT = 9, Internação = mais_que_uma. A conclusão da regra é igual ao valor do atributo meta do registro (Internação = mais_que_uma). Logo, os valores correspondentes deverão ser incrementados na primeira coluna da matriz de confusão, conforme mostra a Figura 3.11. Como o antecedente no registro é 100% igual ao antecedente na regra, então o valor $min = min(\mu_{CON}(1), \mu_{SADT}(9)) = min(1.0, 1.0) = 1.0$ é lançado na primeira linha.

Classe real do exemplo

Classe prevista	***internação = mais_que_uma***	***internação ≠ mais_que_uma***
internação = mais_que_uma	Reg1 + Reg2 = 1.0 + 0.5 = Soma = 1.5	Reg3 + Reg4 = 0.0 + 0.5 = Soma = 0.5
internação ≠ mais_que_uma	Reg1 + Reg2 = 0.0 + 0.5 = Soma = 0.5	Reg3 + Reg4 = 1.0 + 0.5 = Soma = 1.5

Figura 3.11: Exemplo de matriz de confusão difusa

O segundo registro forneceu $min = 0.5$ e a classe da regra também é igual a classe do registro. Logo, se deve lançar 0.5 na posição (1,1) e 0.5 na posição (2,1) da matriz de confusão, ou seja, o complemento de min deve ser lançado na segunda linha da mesma coluna dado que o antecedente no registro difere em 50% (em termos difusos) do antecedente da regra.

No terceiro registro, como a classe é "nenhuma" (diferente de "mais_que_uma") o valor de min deve ser lançado na segunda coluna, mas na segunda linha, pois, o antecedente do registro é completamente diferente do antecedente da regra ($min = 0$).

Um processamento análogo é aplicado ao quarto registro. A soma de todas as colunas corresponde ao número total de registros do conjunto de dados de treinamento, que no exemplo é igual a 4.

Após calcular a matriz de confusão difusa pode-se calcular a aptidão de um indivíduo, no qual foi levada em consideração uma característica importante dos dados utilizados: os pequenos disjuntos[1]. Em geral, algoritmos de indução de regras têm um *bias*[2] que favorece a descoberta de grandes disjuntos ao invés de pequenos disjuntos. Isso se deve à crença de que especializações no conjunto de treinamento são indesejáveis na validação sobre o conjunto de teste ou ao argumento de que os pequenos disjuntos não deveriam ser incluídos no conjunto de regras descobertas uma vez que eles tendem a ser uma das causas de erros na classificação nos dados de teste.

Entretanto, apesar de um pequeno disjunto cobrir geralmente um pequeno número de registros se comparado a todo conjunto de registros, esse pequeno disjunto pode conter um número significativo de registros se todo o conjunto de dados for grande. Esse é o caso dos dados utilizados nesta dissertação, pois tratam-se de 137.020 registros no qual a maioria das regras cobrem de 25 a 500 registros, que correspondem à no máximo 0,5% do total de registros.

As métricas encontradas na literatura para calcular a aptidão em PG não beneficiam a busca por regras pertencentes aos pequenos disjuntos. Bojarczuk *et al.* (2000, 2004) utiliza uma combinação de sensibilidade (Se), especificidade (Sp) e simplicidade (Sy) para calcular a aptidão dos indivíduos, em que:

$$Se = \frac{vp}{vp + fn} \tag{3.10}$$

$$Sp = \frac{vn}{vn + fp} \tag{3.11}$$

$$Sy = \frac{maxnós - 0,5 * numnós - 0,5}{maxnós - 1} \tag{3.12}$$

onde $numnós$ é o número atual de nós (funções e terminais) de um indivíduo, e $maxnós$ é o número máximo de nós permitido na árvore. Mais detalhes dessa métrica pode ser visto em (Bojarczuk *et al.*, 2000, 2004). Assim, o valor da aptidão é calculado como:

$$aptidão = Se * Sp * Sy \tag{3.13}$$

[1]Regras que cobrem um pequeno número de exemplos (Carvalho, 2005)
[2]Inclinação ou tendência

A questão chave nessa métrica é a sensibilidade (Se) e a especificidade (Sp). Observe que nas Equações 3.10 e 3.11 os termos fn e vn são os registros que não são cobertos pela regra. Tome como exemplo, considerando os dados utilizados nesta dissertação, uma regra $R1$ que cubra 150 registros, em que $vp = 140$, $fp = 10$, $vn = 132000$ e $fn = 4870$. Utilizando especificidade por sensibilidade tem-se:

$$Se = \frac{140}{140 + 4870} = 0,0279441 \tag{3.14}$$

$$Sp = \frac{132000}{132000 + 10} = 0,9999242 \tag{3.15}$$

Assim:

$$Se * Sp = 0,0279441 * 0,9999242 = 0,0274389 \tag{3.16}$$

Independente do tamanho da árvore do indivíduo (assumindo $Sy = 1$ que é o valor máximo que esse parâmetro pode assumir quando a regra é tão simples que contém somente um termo) a aptidão desse indivíduo não será maior que 2,75%.

Nesta dissertação, para calcular a qualidade da regra, utilizou-se uma métrica conhecida como precisão (Tan *et al.*, 2006), que visa analisar somente as instâncias cobertas pela regra. A precisão determina a fração de registros que realmente são positivos no grupo que o classificador declarou como classe positiva. Quanto maior a precisão menor o número de falsos positivos classificados pelo classificador. Uma definição formal dessa métrica pode ser vista por meio da Equação 3.17.

$$precisão = \frac{vp}{vp + fp} \tag{3.17}$$

Assim, tome novamente a regra $R1$ com todos os valores definidos anteriormente. Na métrica de precisão tem-se:

$$precisão = \frac{140}{140 + 10} = 0,93333 \tag{3.18}$$

Portanto, observa-se por meio desse exemplo que a métrica de precisão busca não prejudicar as regras pertencentes aos pequenos disjuntos nem dos grandes disjuntos. Assim essa métrica permite uma competição mais justa entre os grandes disjuntos e os pequenos disjuntos no processo evolutivo.

Entretanto, não se pode aplicar diretamente a equação de precisão ao indivíduo, pois um indivíduo pode representar uma ou várias regras, como visto na Seção 3.2.2. Então,

visando beneficiar o indivíduo como um todo, optou-se por determinar a aptidão do indivíduo como sendo a média aritmética simples de todas as regras pertencentes ao indivíduo, conforme a Equação 3.19.

$$média = \frac{\sum_{i=1}^{n} \left(\frac{vp_i}{vp_i + fp_i} \right)}{n} \tag{3.19}$$

onde n é o número de regras pertencentes ao indivíduo.

Em experimentos com essa função de aptidão percebeu-se que a precisão beneficiava regras com cobertura mais baixa do que seria desejável para os objetivos da aplicação escolhida para esta técnica. Para resolver esse problema optou-se por punir indivíduos que possuem regras com menos de 25 casos (definido empiricamente) de *vp* (verdadeiro positivo), nesses casos a regra recebe apenas 5% do valor da precisão calculada originalmente. A equação final para o cálculo da aptidão é exibida pela Equação 3.20:

$$aptidão = \begin{cases} \sum_{i=1}^{n} \left(\frac{vp_i}{vp_i + fp_i} \right) / n, & se \quad vp > 25; \\ \sum_{i=1}^{n} \left(\left(\frac{vp_i}{vp_i + fp_i} \right) * 0.05 \right) / n, & se \quad vp \leq 25. \end{cases} \tag{3.20}$$

Uma das características do algoritmo é que o melhor indivíduo da população deve possuir as melhores regras. Buscando aprimorar essa característica do algoritmo, determinou-se que indivíduos que possuem menos de três regras também devem ser penalizados. O valor dessa punição foi arbitrado como sendo de 50%, ou seja, a aptidão desses indivíduos recebe apenas 50% do valor da aptidão calculada originalmente.

3.2.6 Fluxograma do Sistema

A Figura 3.12 apresenta um fluxograma para entendimento do funcionamento geral do algoritmo. Para o início da execução, os atributos e seus domínios são fornecidos para geração da população inicial. A partir da população inicial cada indivíduo é avaliado com base no conjunto de treinamento e recebe um valor de aptidão, então os melhores indivíduos são selecionados para elitismo. Com base em toda a população os indivíduos são selecionados por torneio procurando os bons indivíduos, que serão submetidos aos operadores genéticos. Os indivíduos resultantes das operações genéticas são inseridos na nova população e o processo se repete até que a nova população esteja completa, então os indivíduos preservados pelo elitismo são inseridos na nova população. Com a nova população pronta os indivíduos sofrem nova avaliação e, caso atinjam o critério de término, seleciona-se o melhor indivíduo, do qual obtém-se as melhores regras que são

apresentadas ao usuário. Caso contrário o processo se repete até que o critério de término seja alcançado.

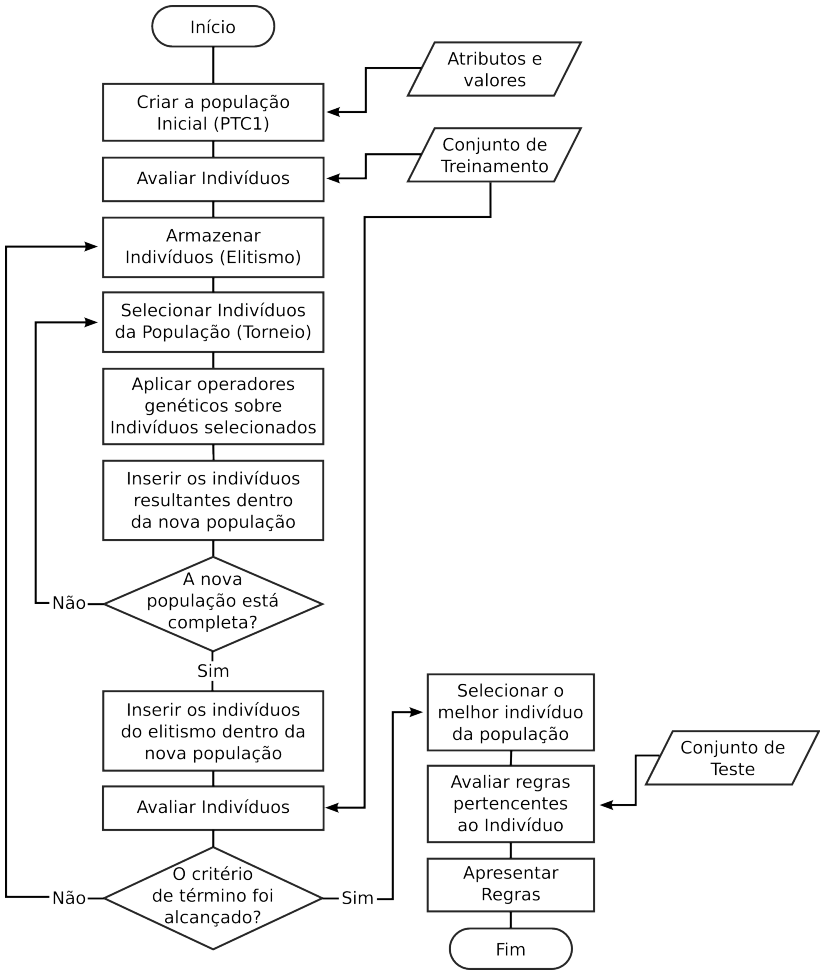

Figura 3.12: Fluxograma do algoritmo PG

3.3 Considerações Finais

A etapa de MD requer a escolha de uma técnica adequada para efetuar a busca de conhecimento oculto em bases dados. A técnica escolhida foi o Algoritmo Cultural (AC) no qual é indispensável para seu funcionamento um algoritmo que utilize população de indivíduos para compor o núcleo do AC. O algoritmo escolhido foi PG, uma vez que AC utilizando PG é pouco explorado no contexto de MD.

Uma característica dos dados utilizados nesta dissertação é que possuem muitas regras pertencentes aos pequenos disjuntos e o algoritmo apresentado neste capítulo leva em

consideração essa característica, propondo uma abordagem do cálculo da aptidão dos indivíduos diferente das encontradas na literatura para PG em MD.

A abordagem apresentada simplifica a etapa de preparação de dados devido ao tratamento dos atributos contínuos utilizando lógica difusa, pois elimina a necessidade de alteração desses atributos na etapa de pré-processamento, além de incorporar algum conhecimento do usuário sobre o domínio de alguns atributos contínuos.

Capítulo

4

Algoritmo Cultural Aplicado à Mineração de Dados

4.1 Considerações Iniciais

Um dos objetivos desta dissertação é o desenvolvimento de uma algoritmo cultural com espaço populacional baseado em programação genética para resolver a tarefa de classificação em MD.

O desenvolvimento desse algoritmo tem o propósito de responder às seguintes questões:

- Qual a influência dos tipos de conhecimento (situacional, domínio e topográfico) utilizados no algoritmo cultural na descoberta de conhecimento?
- Como implementar as impressões gerais do usuário como conhecimentos do AC?
- Qual a sensibilidade do algoritmo cultural em relação aos seus parâmetros iniciais?
- Os resultados obtidos pelo algoritmo proposto são similares aos resultados encontrados por outros algoritmos de MD?

As respostas para essas questões são relevantes para aferir a capacidade de um algoritmo cultural resolver o problema em questão (tarefa de classificação), além de avaliar as vantagens e desvantagens desta abordagem.

O desenvolvimento do algoritmo levou em consideração aspectos como portabilidade, facilidade de modificações/melhorias e desempenho, por isso o algoritmo foi codificado em linguagem *Java* versão 6.0 utilizando o ambiente de desenvolvimento *Eclipse* 3.6.

4.2 Modelo do Sistema

O algoritmo desenvolvido nesta dissertação segue os passos apresentados pelo algoritmo da Figura 4.1. Em seguida é apresentado na Figura 4.2 um fluxograma com mais detalhes do algoritmo para facilitar a compreensão.

```
Algoritmo Cultural
Inicializar a População;
Avaliar a população Inicial;
Criar Espaço de Crenças;
Repita
   Repita
      Selecionar Pais;
      Gerar novos Indivíduos pelas Funções de Influência;
   Até (Nova população estar completa)
   Avaliar os novos Indivíduos;
   Atualizar Espaço de Crenças;
   Atualizar Parâmetros;
Até (Condição de Parada)
```

Figura 4.1: Algoritmo Cultural Implementado

4.2.1 Inicialização, Avaliação da População e Criação do Espaço de Crenças

O primeiro passo na execução é inicializar a população de indivíduos que passará pelo processo de evolução. Como apresentado no Capítulo 3, os indivíduos são representados como árvores e a população é inicializada utilizando o algoritmo PTC1 (ver subseção 3.2.2).

O segundo passo é avaliar a aptidão dos indivíduos da população inicial. Nesta dissertação o valor da função de aptidão é calculado de acordo com a Equação 3.20 da Subseção 3.2.5, podendo sofrer variações que serão apresentadas de acordo com os experimentos do Capítulo 6.

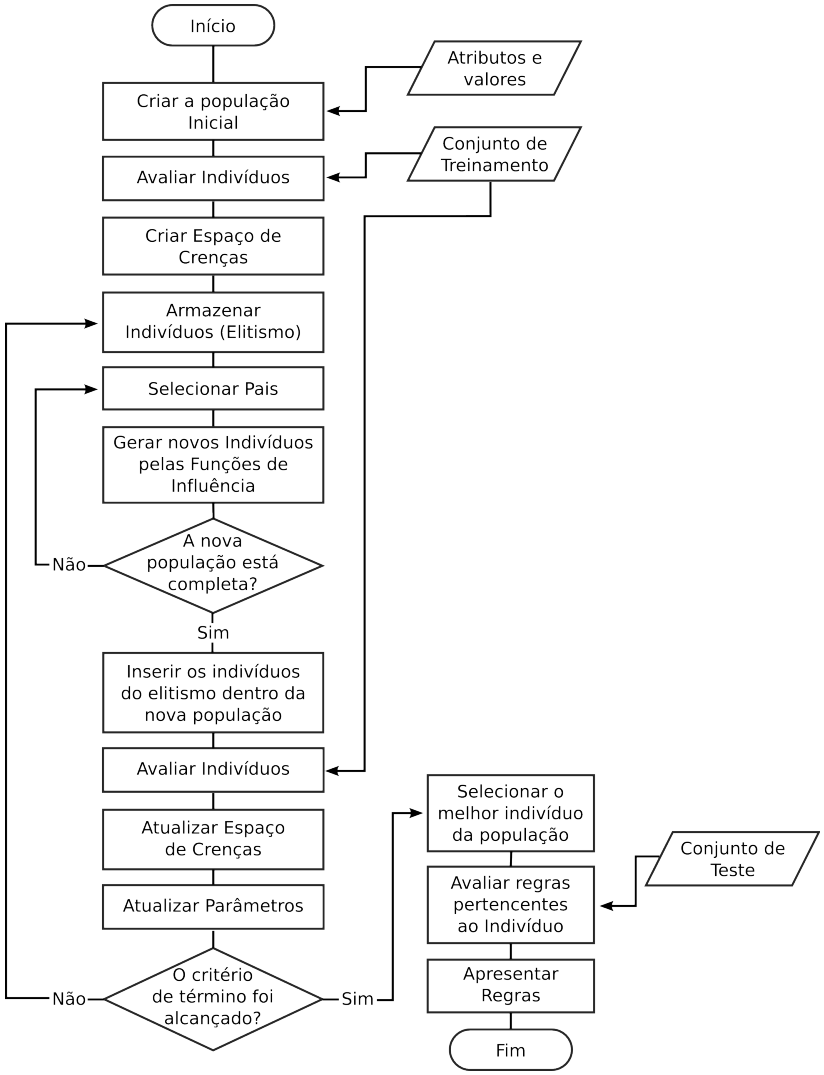

Figura 4.2: Fluxograma do Algoritmo Cultural Desenvolvido

O terceiro passo na execução é criar o espaço de crenças. Essa criação envolve a inicialização dos tipos de conhecimento utilizados pelo algoritmo e das probabilidades de utilização dos mesmos.

Após esses passos o algoritmo entra em um laço de repetição principal em que é executado até que a condição de parada seja alcançada.

4.2.2 Geração e Seleção de Indivíduos e Atualização do Espaço de Crenças

Os filhos são gerados por meio da aplicação das funções de influência (ver subseção 4.5), que são modificações dos operadores de mutação e cruzamento para utilização do conhe-

cimento armazenado ao longo da evolução no espaço de crenças. A avaliação dos filhos é feita da mesma maneira que a avaliação dos indivíduos da população inicial.

Todos os filhos gerados são selecionados para compor a população da próxima geração (método *generational*). Após a nova população estar completa, novos conhecimentos são extraídos e utilizados na atualização do espaço de crenças.

O laço de repetição principal é concluído com alguns cálculos que servem para atualizar os principais parâmetros do algoritmo (ver seção 4.7), os quais serão utilizados na próxima iteração do laço principal. A seguir serão detalhados os principais componentes do algoritmo.

4.3 Espaço Populacional

O espaço populacional de um algoritmo cultural é o componente responsável pela microevolução do algoritmo e geralmente corresponde a algum algoritmo de computação evolutiva. Nesta dissertação o espaço populacional é implementado utilizando PG (como apresentado no Capítulo 3).

Como a grande maioria dos algoritmos de PG, o espaço populacional possui uma população de programas (chamados de indivíduos nesta dissertação por causa da terminologia dos algoritmos culturais), operadores genéticos e um método de seleção dos pais.

Os operadores genéticos utilizados são uma variação do cruzamento e da mutação da PG visto no Capítulo 3. Essa variação foi necessária para a utilização da função de influência e será melhor detalhada juntamente com os tipos de conhecimento. Já a seleção dos pais ocorre por meio de um torneio, como descrito na Subseção 3.2.3. Na implementação utilizada nesta dissertação os filhos gerados pelos operadores genéticos serão os indivíduos que formarão a próxima população e os indivíduos da geração atual não farão parte das gerações futuras, a não ser pela utilização do componente elitista, que visa, neste algoritmo, somente preservar o melhor indivíduo encontrado ao longo das gerações.

4.4 Espaço de crenças

O espaço de crenças é o componente responsável pela macroevolução do algoritmo cultural. A sua função é armazenar os tipos de conhecimentos utilizados pelo algoritmo, conhecimentos esses que guardam as experiências extraídas dos indivíduos ao longo das gerações.

Nesta dissertação foram utilizados três tipos de conhecimentos: o conhecimento situacional, conhecimento topográfico e o conhecimento de domínio. A maneira como foram desenvolvidos é melhor detalhada nas Subseções 4.5.1, 4.5.2 e 4.5.3.

Como descrito na Subseção 2.4.4, uma característica importante do espaço de crenças é a maneira como seus protocolos de comunicação (funções de aceitação e funções de influência) são implementados. Nesta dissertação utilizou-se a função de aceitação dinâmica, em que a quantidade de indivíduos aceitos varia de geração para geração. Já a função de influência escolhe qual dos tipos de conhecimento será utilizado para influenciar a geração dos indivíduos. Para isso a função de influência adapta geração a geração as probabilidades de cada tipo de conhecimento de acordo com o sucesso que eles obtiveram na geração anterior. As demais características dos conhecimentos serão detalhadas a seguir juntamente com seus respectivos conhecimentos.

4.5 Tipos de Conhecimento

Os tipos de conhecimento compõem o espaço de crenças do algoritmo cultural e merecem atenção especial. Como citado, anteriormente foram implementados os conhecimentos situacional, topográfico e de domínio. Esses tipos de conhecimentos foram escolhidos por serem os de mais fácil implementação para resolver o problema desta dissertação.

Cada conhecimento tem diferentes influências nos operadores do espaço populacional. No caso particular do algoritmo proposto, o conhecimento interfere somente no operador de cruzamento, pois em experimentos preliminares houve problemas na utilização da função de influência sobre o operador de mutação. Esses problemas são apresentados na Subseção 4.5.4.

Nas Subseções 4.5.1, 4.5.2 e 4.5.3 são apresentadas as formas como foram desenvolvidos os conhecimentos.

4.5.1 Conhecimento Situacional

O conhecimento situacional tem por característica armazenar um ou mais dos melhores indivíduos encontrados ao longo do processo evolutivo. O objetivo em armazenar esses bons indivíduos é utilizá-los como modelos ou guias na evolução dos demais indivíduos. Esse tipo de conhecimento supõe que bons indivíduos possuem valores para seus genes mais próximos dos valores ótimos e que aproximar os valores dos demais indivíduos da população desses valores é algo benéfico. Uma característica interessante desse conhecimento é que mesmo se um bom indivíduo deixar de existir na população, o

seu conhecimento pode estar preservado no conhecimento situacional, se esse indivíduo pertencer ao conjunto de melhores indivíduos encontrado até aquele momento.

A atualização do conhecimento situacional corresponde a manter sempre atualizada a lista de melhores indivíduos. O número exato de melhores indivíduos armazenados nesse conhecimento depende do problema ao qual está sendo aplicado o algoritmo. A maneira como essa atualização é feita pode ser melhor entendida de acordo com o algoritmo da Figura 4.3.

```
Atualização do Conhecimento Situacional
Se num_melhores < num_max_melhores
    Adiciona Indivíduo
Senão Se indivíduo é melhor que pior_individuo_armazenado
    Remove pior_individuo_armazenado
    Adiciona individuo
Fim Se
```

Figura 4.3: Atualização do Conhecimento Situacional

Na Figura 4.3 *num_melhores* é o número de melhores indivíduos armazenados até o momento, *num_max_melhores* é o número máximo de indivíduos que o conhecimento armazena (definido como parâmetro inicial), *individuo* é o indivíduo candidato a participar do conjunto de melhores indivíduos e *pior_individuo_armazenado* é o pior indivíduo do conjunto de melhores indivíduos armazenados.

A maneira como esses indivíduos armazenados influenciarão as próximas gerações é descrito a seguir.

Função de Influência do Conhecimento Situacional no Cruzamento

No cruzamento influenciado pelo conhecimento situacional um dos pais utilizados na geração dos filhos é escolhido pelo processo de seleção descrito na Subseção 2.3.5 (seleção por torneio a partir de toda população) e o outro pai é um dos indivíduos armazenados no conhecimento situacional.

O cruzamento foi implementado como uma combinação dos dois indivíduos no qual gera-se dois filhos. A combinação permite que os filhos gerados estejam a uma distância intermediária dos dois pais, ou seja, permite explorar a região do espaço de busca entre os pais e, portanto, como um dos pais é um dos melhores indivíduos encontrados ao longo da evolução, esse operador ajuda a explorar bem as regiões do espaço de busca ao redor dos melhores indivíduos. A Figura 4.4 exemplifica melhor o processo.

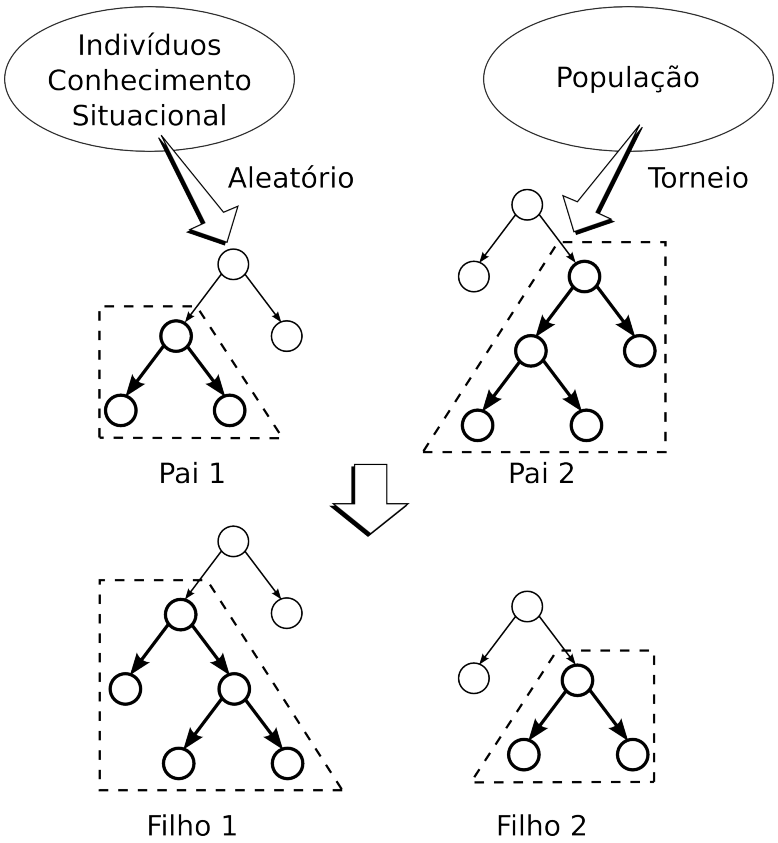

Figura 4.4: Influência do Conhecimento Situacional no cruzamento

Contudo, como citado no Capítulo 3, o cruzamento em árvores possui uma característica destrutiva, podendo gerar indivíduos inválidos, (ver subseção 3.2.2). Assim, como na PG utilizou-se a técnica *Brood Recombination* no cruzamento de todos os conhecimentos implementados do algoritmo cultural.

4.5.2 Conhecimento Topográfico

O propósito do conhecimento topográfico é criar um mapa do espaço de busca durante o processo evolucionário, em que se utiliza algumas características importantes do problema em questão e que se queira levar em consideração na resolução do problema. A partir dessas características pode-se criar um mapa com uma amostragem dos indivíduos que se enquadram nessas características. Por exemplo, nesta dissertação as características consideradas importantes são: interesse do usuário e taxa de acerto de uma regra. Como, segundo a abordagem utilizada nesta dissertação, um indivíduo pode possuir um conjunto de regras, criou-se um mapa no qual cada região contém uma amostra dos indivíduos que possuem sua melhor regra com essas características.

As regiões são denominadas células e são ordenadas pela média dos valores de aptidão dos indivíduos pertencentes a célula. Como a aptidão de um indivíduo não é apenas a taxa de acerto e interesse de uma regra, mas sim a média da taxa de acerto de um conjunto de regras pertencente ao indivíduo, além de outras características como cobertura mínima e número de regras por indivíduo o conhecimento topográfico tenta agrupar essas características buscando melhorar a qualidade dos indivíduos. Por exemplo, considere um indivíduo que possua uma regra com interesse e taxa de acerto altos e outras duas regras com interesse e taxa de acerto baixos, consequentemente esse indivíduo possuirá uma aptidão baixa devido as duas regras com interesse e taxa de acerto baixos. Assim, as chances desse indivíduo ser um representante na célula é pequeno. Agora, se o indivíduo possuir uma regra com interesse e taxa de acerto alto e duas com taxa de acerto e interesse médios terá uma aptidão melhor e consequentemente uma maior chance de ser escolhido como representante da célula. Portanto, o conhecimento topográfico tenta agrupar os indivíduos que possuam características que, de alguma maneira, maximize a aptidão e, como citado no exemplo anterior, são os indivíduos que possuem boas regras, boa cobertura e um bom número de regras.

A Figura 4.5 exibe um exemplo de um mapa de espaço de busca para *interesse* x *taxa de acerto*.

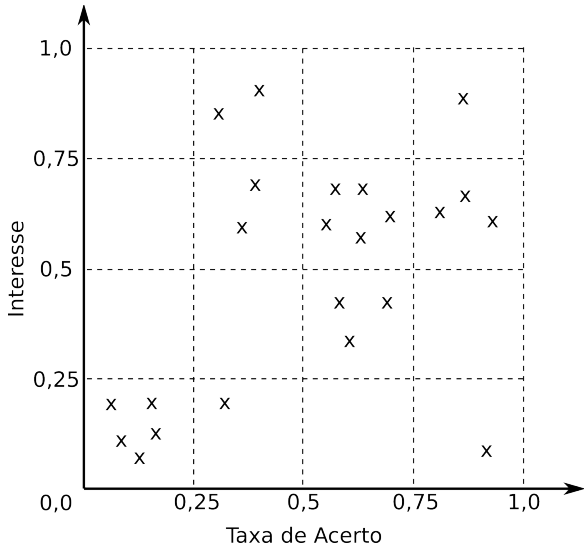

Figura 4.5: Exemplo de mapa para o Conhecimento Topográfico

Na abordagem utilizada neste trabalho, o tamanho das células possuem valores determinados como os da Figura 4.5 e a atualização do conhecimento topográfico corresponde a manter sempre atualizadas as células com os melhores indivíduos, ou seja, se um indivíduo possui valor de aptidão melhor para uma dada célula do que algum indivíduo pertencente

a célula, então essa célula é atualizada, desde que esse indivíduo respeite os limites da célula. Utilizou-se nesta abordagem um número máximo de indivíduos por célula com o objetivo de economizar memória e melhorar o desempenho. Na atualização, o pior indivíduo é substituído pelo novo indivíduo. A maneira como essa atualização é feita pode ser melhor entendida com o algoritmo da Figura 4.6.

```
Atualização do conhecimento topográfico
Para (Cada célula)
    Se (Indivíduo pertence aos limites da célula)
        Se (num_indivíduos >= max_indivíduos)
            Se (aptidão_indivíduo > aptidão_pior_indivíduo)
                Remove pior indivíduo da célula
                Adiciona indivíduo
        Senão
            Adiciona indivíduo
        Fim Se
    Fim Se
Fim Para
```

Figura 4.6: Atualização do Conhecimento Topográfico

Na Figura 4.6 *individuo* é o indivíduo candidato a participar da célula, *num_individuos* é o número atual de indivíduos da célula, *max_individuos* é o número máximo de indivíduos permitidos em uma célula e *aptidao_individuo* e *apitdao_pior_individuo* são a aptidão do indivíduo candidato e do pior indivíduo da célula respectivamente.

Função de Influência do Conhecimento Topográfico no Cruzamento

No cruzamento influenciado pelo conhecimento topográfico um dos pais utilizados na geração dos filhos é escolhido a partir de toda a população utilizando torneio, como descrito anteriormente. O outro pai é escolhido de maneira aleatória a partir das n melhores células. Cabe ressaltar que a qualidade de uma célula é medida de acordo com a média de aptidão dos indivíduos pertencentes a ela. A Figura 4.7 exemplifica o processo de cruzamento.

Com esse cruzamento busca-se levar os indivíduos para regiões mais promissoras. A Figura 4.8 ilustra essa característica.

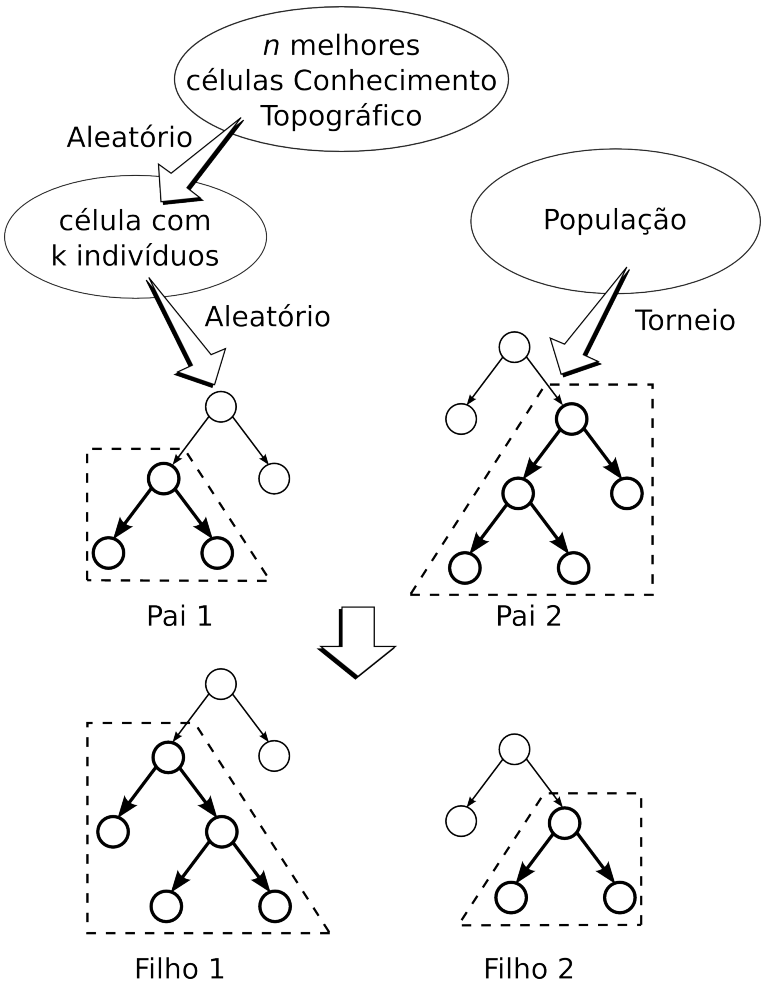

Figura 4.7: Influência do Conhecimento Topográfico no cruzamento

4.5.3 Conhecimento De Domínio

O conhecimento de domínio, como o próprio nome diz, é referente ao domínio do problema em questão. Para um dado problema é possível que se conheça algo a respeito da forma ou topologia do espaço de busca. Nesta dissertação esse conhecimento representa as crenças do usuário, caracterizando um método *user-driven* de medida de interesse.

As crenças do usuário utilizadas são baseadas em impressões gerais (IG's), que são um tipo específico de conhecimento prévio no qual o usuário tem a respeito do domínio do problema.

Para utilizar as IG's juntamente com as regras é necessário que elas possuam a mesma sintaxe e semântica, pois é necessário realizar a influência desse conhecimento sobre a população.

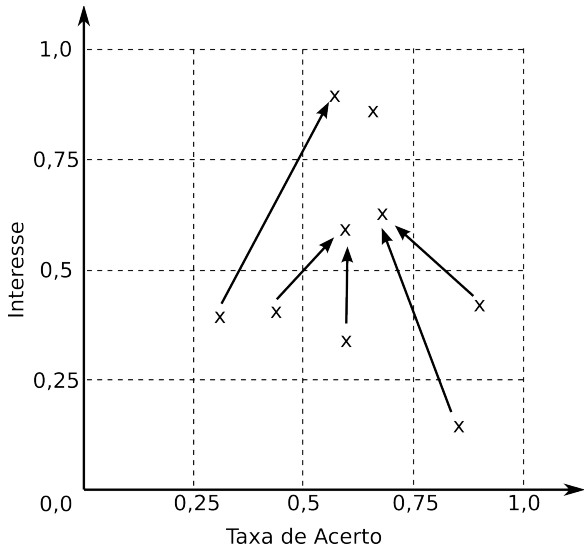

Figura 4.8: Influência do Conhecimento Topográfico

A atualização do conhecimento de domínio não é realizada pelo algoritmo, uma vez que as informações desse conhecimento dependem do usuário, ou seja, de fora do sistema. Trabalhos futuros podem avaliar a viabilidade e a importância em incorporar também técnicas *data-driven* nesse conhecimento, permitindo assim a atualização automática do conhecimento.

A influência direta das IG's sobre a população é um grande diferencial dessa técnica em relação as demais, que geralmente utilizam as IG's de maneira indireta para o cálculo da aptidão ou no pós-processamento. Esse conhecimento permite utilizar as impressões gerais diretamente no processo evolutivo, por meio da função de influência, e a maneira como isso é feito é apresentado a seguir.

Função de Influência do Conhecimento de Domínio no Cruzamento

A influência desse conhecimento no cruzamento é realizada da mesma maneira que o situacional e para isso é necessário transformar as IG's em indivíduos para permitir o cruzamento com os indivíduos da população.

Para transformar as IG's em indivíduos considere, por exemplo, que o usuário forneceu n IG's. Dessas n IG's, três contradizem o atributo meta, ou seja, o atributo meta é igual ao do algoritmo, mas o valor é diferente. Então essas três IG's vão fazer parte do conhecimento de domínio e são transformadas em árvores utilizando somente o operador E como função. A Figura 4.9 exemplifica os indivíduos gerados utilizando IG's.

Perceba que os indivíduos gerados a partir das IG's contém somente uma regra, pois em experimentos com indivíduos com múltiplas IG's (regras) ocorria muitos casos em que

no cruzamento uma IG completa (todos os antecedentes de uma IG) passava para o filho, o que não é desejável, pois o interesse é em somente parte de uma IG (ver Seção 4.6).

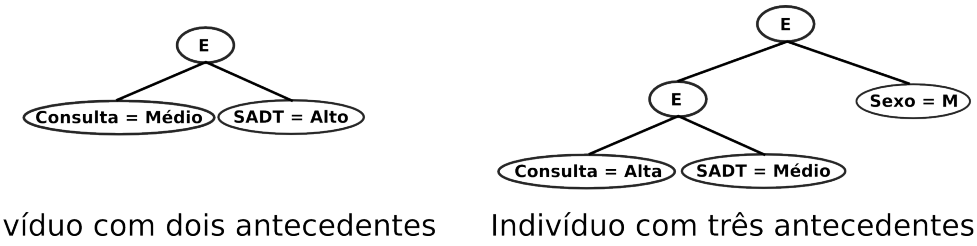

Figura 4.9: Indivíduos gerados a partir de IG's

Os indivíduos gerados por meio das IG's passam então a fazer parte do conhecimento de domínio e influenciam no cruzamento conforme a Figura 4.10.

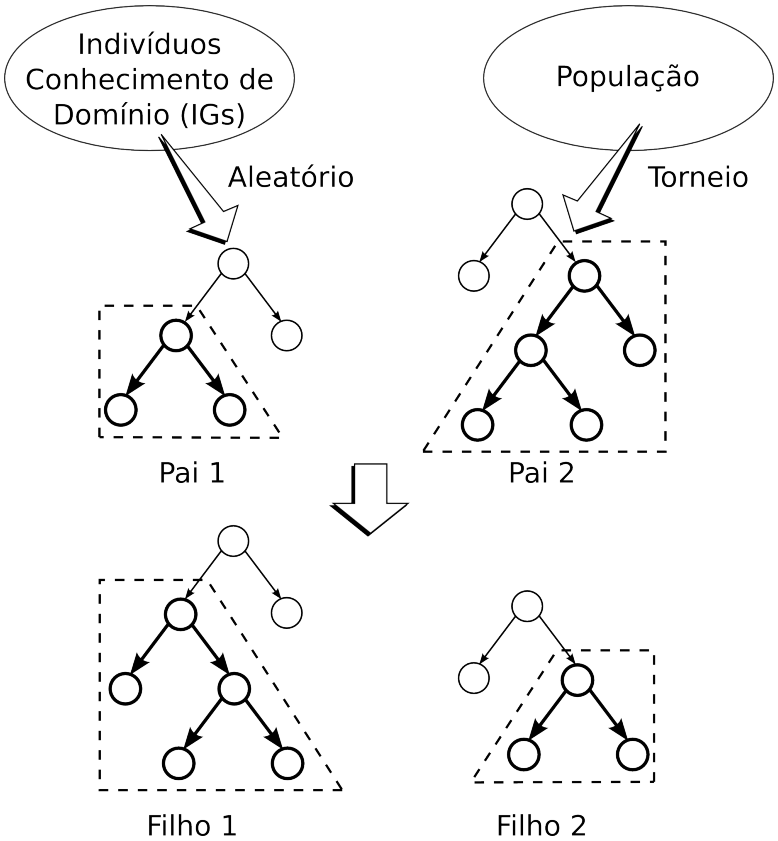

Figura 4.10: Influência do Conhecimento de Domínio no cruzamento

4.5.4 O problema da mutação nos conhecimentos

A princípio tentou-se utilizar as funções de influência na mutação para cada um dos três conhecimentos utilizados nesta dissertação. A mutação consistia basicamente em escolher

aleatoriamente um nó terminal dos indivíduos do conhecimento utilizado e aplicar também de maneira aleatória nos filhos gerados. Essa abordagem mostrou-se ruim, pois tornava o espaço de busca restrito aos terminais (atributo e valor) utilizados na população inicial e nos conhecimentos, evitando assim que outros terminais fossem utilizados no processo evolutivo.

Tentou-se encontrar outras maneiras de se utilizar a influência dos conhecimentos na mutação, entretanto, não se encontrou uma abordagem satisfatória para esse problema. Trabalhos futuros podem tentar encontrar meios de se aplicar as funções de influência na mutação. Nesta dissertação a mutação é aplicada da mesma maneira que em PG (ver Subseção 3.2.3), sem a utilização de funções de influência.

4.6 Interesse nas Regras

Nesta dissertação constatou-se duas maneiras de se utilizar o interesse em regras para algoritmos evolutivos. A primeira utiliza o cálculo do interesse na aptidão durante a evolução, em que um mesmo algoritmo realiza a busca por regras de previsão, por meio de um processo evolutivo e ao mesmo tempo avalia cada indivíduo gerado quanto ao interesse na regra por ele representado (além da avaliação da qualidade da regra), permitindo a evolução preferencial de indivíduos que atendam ao interesse do usuário. A segunda é após o processo evolutivo, em que calcula-se o interesse das regras da população final e apresentam-se as melhores.

Independentemente de quando se avaliar o interesse é necessário fazer o cálculo do grau de interesse nas regras, e para isso utilizou-se a mesma abordagem de Romão (2002), em que o grau de similaridade é calculado como a diferença entre cada regra R comparada com cada impressão geral existente no conjunto de IG's. Por esse motivo, as regras em IG e em R devem possuir a mesma sintaxe e semântica. Cada regra gerada pelo algoritmo é chamada R_i, enquanto que cada impressão geral (conhecimento prévio) é identificado por IG_j, conforme Figura 2.6 do Capítulo 2.

Como exemplo considere as seguintes regras de previsão descobertas:

R_1: (Grupo_CID_V = Sim), (SADT = Alto) → Consulta = Medio

R_2: (Internação = Mais_Que_Uma), (Grupo_Cid_XII = Sim), (SADT = Alto) → Consulta = Alto

R_3: (Grupo_CID_XII = Sim), (SADT = Alto) → Consulta = Baixo

Assumindo que o usuário acredita que os beneficiários que utilizaram muito o SADT e tiveram mais que uma internação em geral precisam de muitas consultas, esta impressão geral (hipótese) pode ser representada pela seguinte IG difusa:

IG_j: (Internação = Mais_Que_Uma), (SADT = Alto) → Consulta = Alto

O significado do termo difuso "alto" deve ser fornecido pelo usuário (ver subseção 3.2.1).

Dependendo do propósito do usuário, pode-se implementar algoritmos diferentes. Por exemplo, para a confirmação de hipóteses, deve-se considerar as regras mais similares às IG's como sendo as melhores. Retornando o exemplo anterior e considerando a similaridade com IG_j, as regras descobertas podem aparecer na seguinte ordem decrescente de preferência:

R_2: (Internação = Mais_Que_Uma), (Grupo_CID_XII = Sim), (SADT = Alto) → Consulta = Alto

R_1: (Grupo_CID_V = Sim), (SADT = Alto) → Consulta = Medio

R_3: (Grupo_CID_XII = Sim), (SADT = Alto) → Consulta = Baixo

Nesse caso, a regra R_2 é bastante semelhante à IG (contendo dois atributos previsores iguais), facilitando a sua escolha baseada no grau de similaridade. A regra R_1 é preferida à R_3 devido à quantidade de consultas ser mais próximo (Médio > Baixo).

Por outro lado, se o propósito é encontrar regras que contradizem as hipóteses, a ordenação resultante deverá ser:

R_3: (Grupo_CID_XII = Sim), (SADT = Alto) → Consulta = Baixo

R_1: (Grupo_CID_V = Sim), (SADT = Alto) → Consulta = Medio

R_2: (Internação = Mais_Que_Uma), (Grupo_CID_XII = Sim), (SADT = Alto) → Consulta = Alto

Nesse caso, a regra R_3 contradiz bastante a hipótese de consulta alta, facilitando sua escolha. Em caso de empate a escolha é viabilizada pelo valor do grau de interesse. O cálculo do grau de interesse será visto na subseção 4.6.2.

Tanto a abordagem no qual busca-se regras similares as IG's, quanto as que contradizem exigem a quantificação do grau de similaridade do antecedente das regras, assunto da subseção 4.6.1. Nesta dissertação optou-se pela segunda abordagem, buscando por regras que contradizem as IG's, com o propósito de descobrir conhecimento novo e surpreendente para o usuário.

4.6.1 Cálculo da Similaridade do Antecedente da Regra

Independentemente do tipo de interesse do usuário (se está interessado em regras similares às IG's ou que contradizem às IG's), é importante calcular o grau de similaridade do antecedente, que é utilizado em ambos os casos.

Para calcular a similaridade do antecedente (SA) da regra (comparado com cada IG) adaptou-se a abordagem de Romão (2002):

$$SA_{(i,j)} = \frac{|A_{i,j}|}{max(|R_i|, |IG_j|)} \tag{4.1}$$

em que: $|R_i|$ é o número de atributos no antecedente da regra descoberta R_i;

IG_j é o número de atributos no antecedente da IG_j;

$A_{(i,j)}$ é o número de atributos de R_i que são iguais (atributo e valor) aos atributos da IG_j.

A Equação 4.1 denota o grau de similaridade do antecedente completo de uma regra (considerando nome e valor dos atributos) e não apenas a similaridade dos nomes dos atributos.

Como exemplo, considere o par R_i (regra) e IG_j (impressão geral):

$|R_i|$: (Especialidade-09 = Sim), (SADT = Baixo) → Consulta = Baixo

IG_j: (Grupo_CID_VIA = Sim), (SADT = Baixo) → Consulta = Baixo

O cálculo da similaridade do antecedente, no exemplo acima aplicado a Equação 4.1 fornece:

$$SA_{(i,j)} = \frac{1}{max(2,2)} = \frac{1}{2} = 0,5 \tag{4.2}$$

O numerador de $SA_{i,j}$ representa a cardinalidade da interseção de atributos no antecedente de R_i e IG_j com o mesmo nome/valor (SADT = 'Baixo'). O denominador representa o número de atributos no antecedente do R_i e IG_j respectivamente, extraindo o máximo deles. O resultado obtido por SA indica que há 50% de similaridade entre os antecedentes de R_i e IG_j.

4.6.2 Cálculo do Grau de Interesse

Quando um usuário está interessado em regras diferentes das IG's, há três situações possíveis, segundo Liu e Hsu (1996):

- Antecedente inesperado;
- Antecedente contraditório;
- Consequente contraditório.

Nesta dissertação utilizou-se a abordagem de consequente contraditório, em que o atributo meta do consequente deve ser o mesmo na IG e na regra, mas o valor do atributo meta deve ser diferente. Nesta abordagem são consideradas as regras em que o antecedente da IG contém pelo menos uma condição igual no antecedente da regra (tanto nome quanto

valor), mas o valor do consequente difere de maneira contraditória. Considerou-se que esse caso possui maior potencial para a descoberta de regras altamente surpreendentes para o usuário, mas os demais casos deverão ser considerados em pesquisas futuras.

O cálculo do grau de interesse de uma regra R_i é obtido comparando o consequente da regra com o consequente das IG's. Se o consequente da regra (atributo e valor) for igual ao consequente da IG, a regra é considerada irrelevante com relação a IG (grau de interesse = 0). Se o nome do atributo do consequente da regra for igual ao nome do atributo do consequente de uma IG mas o valor for diferente, o interesse nesta regra será maior que zero e será determinado da seguinte maneira: se a diferença entre os valores for de "Baixo" para "Alto", então o interesse é máximo, ou seja, equivalente a similaridade do antecedente. Em caso contrário, se um dos valores for igual a "Médio" (diferença entre "Baixo" e "Médio" ou entre "Alto" e "Médio"), então o interesse nessa regra será reduzido em 50% e, para atributos categóricos, o interesse é máximo quando o valor é diferente do atributo meta e zero quando é igual.

Em seguida, pode-se calcular o interesse em uma regra nova específica R_i dado um conjunto de impressões gerais IG, conforme segue:

$$Interesse = Max(SA_{(i,1)}, SA_{(i,2)}, ..., SA_{(i,j)}, ..., SA_{(i,|IG|)},) \tag{4.3}$$

em que $|IG|$ é o número de impressões gerais especificadas pelo usuário.

Cabe destacar que uma regra pode ser considerada relevante mesmo se apenas algumas condições forem equivalentes a uma IG e a conclusão for diferente.

4.6.3 Cálculo da aptidão do Indivíduo

Para cada regra do indivíduo o interesse é calculado e multiplicado pela precisão da regra. Entretanto, só a multiplicação entre interesse e precisão tende a eliminar as regras com interesse 0 pois o produto de $precisão * interesse$ resultaria em zero, impedindo a evolução de regras sem interesse. Buscando solucionar esse problema criou-se uma nova função para o cálculo da qualidade da regra baseado na proposta de Romão (2002), conforme segue:

$$qualidade = \begin{cases} interesse * precisão, & se \quad interesse > 0; \\ precisão/20, & se \quad interesse \leq 0. \end{cases} \tag{4.4}$$

Segundo Romão (2002), quando a ordem de grandeza da precisão é 1/20 (ou menor) a Equação 4.4 para o cálculo da *qualidade* da regra fornece um resultado melhor sem afetar os demais resultados. Observe que multiplicar por 0.05 não muda os resultados uma vez que não tem-se $interesse = 0.05$ (pois é permitido em uma regra no máximo 5

antecedentes, ver Equação 4.1), ou seja, não há possibilidade de conflito entre uma regra com interesse e outra sem interesse com a mesma qualidade.

Assim, considerando a função de qualidade de uma regra apresentada pela Equação 4.4, tem-se a função de aptidão para AC conforme a Equação 4.5.

$$aptidão = \begin{cases} \sum_{i=1}^{n} (qualidade)/n, & se \quad vp > 25; \\ \sum_{i=1}^{n} ((qualidade) * 0.05)/n, & se \quad vp \leq 25. \end{cases} \tag{4.5}$$

em que $vp > 25$ é utilizado para garantir a cobertura mínima assim como em PG (ver Equação 3.20).

Considerando ainda a característica importante da PG que é utilizada como componente populacional do AC, determinou-se que indivíduos que possuem apenas uma regra devem ser penalizados em 50%.

4.7 Adaptação dos Parâmetros

O ajuste de parâmetros em algoritmos evolutivos é uma tarefa complexa e bastante controversa, visto que é muito difícil provar que um determinado conjunto de parâmetros é ótimo para um determinado problema, ainda mais em se tratando de mineração de dados, em que conforme se altera os dados utilizados na mineração dificilmente os parâmetros do algoritmo serão os mesmos para se obter boas soluções.

O controle automático de parâmetros surge como uma alternativa elegante e eficiente para esse problema, e pode ser classificado de acordo com o tipo e nível de adaptação (Eiben *et al.*, 1999).

Segundo Eiben *et al.* (1999) os métodos para alterar o valor de um parâmetro podem ser classificados em três categorias.

- **Controle de parâmetro determinístico**: ocorre quando o valor de um parâmetro estratégico é alterado por alguma regra determinística.
- **Controle de parâmetro adaptativo**: ocorre quando há algum *feedback* da busca que é utilizada para determinar a direção e/ou magnitude da mudança para o parâmetro estratégico.
- **Controle de parâmetro auto-adaptativo**: ocorre quando um parâmetro é modificado como parte do processo evolutivo, ou seja, quando o parâmetro é alterado por mutação, cruzamento ou equivalente. Esse tipo de controle de parâmetros também é denominado Co-Evolução.

Pode-se também definir em que nível dentro dos algoritmos evolucionários a adaptação acontece (Hinterding *et al.*, 1997):

- **Ambiental**: é realizado quando o ambiente ao redor dos indivíduos é modificado;
- **Populacional**: ocorre quando a modificação no parâmetro afeta toda a população;
- **Individual**: ocorre quando um parâmetro é ajustado para um indivíduo em particular;
- **Componente**: ocorre quando a modificação afeta apenas um componente do indivíduo.

Nesta dissertação há vários pontos de controle automático: a função de aceitação, a taxa de mutação, a taxa de cruzamento, probabilidade da influência ser exercida pelos conhecimentos situacional, topográfico ou de domínio.

A taxa de aceitação é o parâmetro que diz quantos indivíduos serão aceitos para contribuir com seu conhecimento no espaço de crenças. O controle desse parâmetro pode ser classificado como determinístico e populacional. Ele é feito pela Equação 4.6.

$$aceitos = taxa_aceitos + (taxa_aceitos/geracao) \tag{4.6}$$

em que *aceitos* é o percentual de indivíduos a serem aceitos para atualização do espaço de crenças, *taxa_aceitos* é um parâmetro de entrada que determina o percentual mínimo de indivíduos aceitos e *geracao* é o número de geração corrente. Essa função aceita muitos indivíduos no início da evolução e progressivamente aceita menos indivíduos conforme o processo evolui.

O controle da taxa de mutação e cruzamento são independentes, visto que na implementação do presente trabalho a mutação é aplicada após o cruzamento com uma certa probabilidade. Ambos são do tipo adaptativo populacional. A atualização dessas taxas são dadas pelas Equações 4.7 e 4.8.

$$\begin{aligned} taxaMutacao = 0,01 + 0,8 * (0,7 * taxaMutacao + \\ 0,3 * (sucessoMutacao/numOperadorMutacao)); \end{aligned} \tag{4.7}$$

$$\begin{aligned} taxaCruzmento = 0,1 + 0,8 * (0,7 * taxaCruzamento + \\ 0,3 * (sucessoCruzamento/numOperadorCruzamento)); \end{aligned} \tag{4.8}$$

em que $taxaMutacao$ e $taxaCruzamento$ correspondem a taxa de mutação e taxa de cruzamento e $sucessoMutacao$ e $sucessoCruzamento$ são a quantidade de sucessos obtidos pelos operadores de mutação e cruzamento respectivamente, e esse sucesso é medido pela melhora que os filhos tiveram em relação aos seus pais, e $numOperadorMutacao$ e $numOperadorCruzamento$ é o número de vezes que o operador de mutação e cruzamento foram acionados.

Essas equações determinam que a taxa mínima para o operador de cruzamento é 10% (constante $0,1$) e de mutação 1% (constante $0,01$) e que a taxa a ser aplicada na próxima geração é uma combinação linear da taxa aplicada na última geração e do sucesso que o operador obteve quando comparado com todas as utilizações do operador na geração atual. A constante $0,8$ nas Equações 4.7 e 4.8 garante que a contribuição máxima dos cruzamentos, mutações e seus sucessos seja no máximo de 80%, o mesmo ocorre com as demais constantes em que $0,7$ permite que 70% do valor da taxa de mutação ou cruzamento anteriores sejam utilizados na próxima geração e $0,3$ permite que sucessos das mutações e cruzamentos contribuam em no máximo 30% com atualização das probabilidade dentro da escala maior de 80%. Essa técnica permite que o algoritmo encontre de maneira eficiente valores adequados para essas taxas.

A probabilidade de se aplicar cada tipo de conhecimento na influência dos operadores também é um tipo de controle de parâmetros adaptativo e populacional e possui regras semelhantes às utilizadas para adaptar as taxas de mutação e cruzamento. As Equações 4.9, 4.10 e 4.11 formalizam a adaptação da probabilidade de se utilizar os conhecimentos situacional, topográfico e de domínio.

$$probSituacional = 0,1 + 0,7 * sucessoSituacional/sucessoConhecimentos \quad (4.9)$$

$$probTopografico = 0,1 + 0,7 * sucessoTopografico/sucessoConhecimentos \quad (4.10)$$

$$probDominio = 0,1 + 0,7 * sucessoDominio/sucessoConhecimentos \quad (4.11)$$

em que $probSituacional$, $probTopografico$ e $probDominio$ são as probabilidades de aplicação das influências situacional, topográfico e de domínio respectivamente e $sucessoSituacional$, $sucessoTopografico$ e $sucessoDominio$ correspondem aos sucessos obtidos

pelos conhecimentos situacional, topográfico e de domínio, ou seja, essas variáveis são *incrementadas em 1(um)* sempre que esses conhecimentos são aplicados aos indivíduos e esses indivíduos são selecionados para fazer parte da população e $sucessoConhecimentos$ é a soma das variáveis $sucessoSituacional$, $sucessoTopografico$ e $sucessoDominio$.

Assim como ocorre com as taxas de mutação e cruzamento, essas equações garantem que a probabilidade mínima de qualquer tipo de conhecimento é de 10% mas diferem das taxas de mutação e cruzamento, pois as probabilidades dos conhecimentos dependem única e exclusivamente do seu desempenho na última geração, que contribuem com 70% do valor da probabilidade (constante 0,7). Isso se deve ao fato de os conhecimentos serem constantemente adaptados ao longo do processo evolutivo, fazendo com que seus efeitos sejam alterados rapidamente, o que exige uma fórmula que permita mudanças rápidas nas probabilidades.

Como pode ser evidenciado nesta seção, o algoritmo proposto é capaz de controlar automaticamente quase todos os seus parâmetros. Os únicos parâmetros que precisam ser determinados a priori são o tamanho da população, número de gerações e quantidade de indivíduos para elitismo.

4.8 Considerações Finais

Este capítulo apresentou a abordagem utilizada para a implementação do algoritmo cultural para resolver a tarefa de classificação em mineração de dados, utilizando como componente populacional um algoritmo de programação genética. Na implementação dos conhecimentos procurou-se aprimorar o processo de busca por regras, no qual um desses aprimoramentos pode ser atribuído ao conhecimento de domínio que utiliza uma abordagem direta de interferência das IG's sobre o processo evolutivo.

Outra característica importante abordada neste capítulo é o controle automático de parâmetros, que se mostra de grande importância, principalmente para mineração de dados, no qual os parâmetros do algoritmo são fortemente dependentes dos dados utilizados para mineração.

Capítulo

5

Preparação de Dados

5.1 Considerações Iniciais

As operadoras de planos de saúde suplementar possuem uma grande quantidade de dados armazenados sobre procedimentos realizados pelos beneficiários. Essas bases de dados podem conter conhecimentos ocultos de alta qualidade que poderão auxiliar na tomada de decisões (e.g.: programas de prevenção de doenças).

Para descoberta de conhecimento oculto em banco de dados é necessário a aplicação de algoritmos de mineração de dados (MD), parte do processo de descoberta de conhecimento em banco de dados (DCBD). No entanto, não importa o quão "inteligente"um algoritmo de MD seja, ele falhará na descoberta de conhecimento de alta qualidade se for aplicado em dados de baixa qualidade (Freitas, 2002b). Como a maioria das bases de dados apresenta diversos problemas (como valores faltando, valores inconsistentes, falta de precisão, ruídos e erros de medição) que reduzem a sua qualidade, há necessidade de se aplicar técnicas de pré-processamento para melhoria da qualidade desses dados.

Considerando os diversos problemas normalmente encontrados em bases de dados como ruídos, atributos duplicados, estrutura diferente da utiliza pelo algoritmo minerador entre outros, aplicou-se técnicas de pré-processamento em dados de um plano de saúde suplementar (PSS), objetivando prepará-los para um estudo de caso dos algoritmos apresentados nos Capítulos 3 e 4. Esses dados contém um total de 159.883 beneficiários. No decorrer das etapas do pré-processamento alguns problemas foram encontrados, como dados armazenados em banco de dados relacional. A maioria dos algoritmos de MD não

conseguem trabalhar com dados armazenados dessa forma, e para resolver esse problema foi necessário analisar esses dados e identificar as tabelas e os atributos dessas tabelas para que pudessem ser utilizados no processo de DCBD.

Segundo Fayyad *et al.* (1996a), a DCBD consiste em nove etapas, mas o presente capítulo é fundamentado nas quatro primeiras: seleção de um conjunto de dados alvo, limpeza de dados, redução e projeção de dados.

Neste capítulo, apresenta-se uma metodologia para aplicação dessas etapas sobre dados de um PSS objetivando tanto demonstrar a metodologia quanto obter dados preparados para serem utilizados pelo algoritmo de MD desenvolvido nesta dissertação e pelo grupo de pesquisa GPEA (Grupo de Pesquisa em Engenharia de Algoritmo) (Romão, 2011) da UEM (Universidade Estadual de Maringá).

5.2 Métodos

Os métodos utilizados neste capítulo baseiam-se nas quatro primeiras etapas do processo de DCBD proposto por Fayyad *et al.* (1996a) com algumas adaptações. Os passos básicos dos métodos aplicados consistem em:

- Entendimento do domínio da aplicação que implica no levantamento do conhecimento prévio e dos objetivos do usuário;
- Seleção de um conjunto de dados alvo, ou seja, selecionar um conjunto de dados ou concentrar-se em um subconjunto ou instâncias sobre os quais a descoberta será efetuada;
- Limpeza de dados, que consiste em operações básicas tais como remoção de dados errôneos e manipulação de atributos com valores ausentes;
- Pré-seleção de atributos relevantes com base no domínio da aplicação e nos objetivos do usuário;
- Redução e projeção de dados, envolve métodos de transformação para reduzir o número efetivo de atributos.

Esses passos e métodos são aplicados e descritos mais detalhadamente nas próximas seções, levando em consideração o objetivo de preparar dados de um PSS.

5.2.1 Entendimento do Domínio da Aplicação

Apesar do presente trabalho abordar apenas a preparação de um conjunto de dados para a aplicação de algoritmos de MD, o objetivo final desta pesquisa, considerando o processo completo de DCBD, é encontrar padrões válidos, compreensíveis e úteis para gestores de planos de saúde suplementar. Para encontrar esses padrões é necessário definir o escopo do processo de DCBD. Nesta pesquisa focou-se em procedimentos, especialidades e diagnósticos de doenças de beneficiários de um plano de saúde.

Os dados foram obtidos exclusivamente para fins de pesquisa, sem qualquer identificação pessoal, a partir de uma base de teste de uma operadora de saúde suplementar contendo informações administrativas e de procedimentos em hospitais, laboratórios e consultórios relativas a beneficiários de um plano de saúde do estado de Santa Catarina, disponibilizados na forma de um banco de dados em *SQL Server*.

É importante salientar que as informações pessoais, relativas aos beneficiários desse plano de saúde, foram excluídas da base, preservando a privacidade dos beneficiários. Além disso, os resultados representam uma reorganização de dados já existentes, não apresentando descrição de perfis ou sugestão de diagnósticos, excluindo a necessidade de aprovação do comitê de ética. Essa exclusão de informações pessoais não causa nenhum prejuízo ao processo de DCBD, visto que o objetivo final da pesquisa é buscar padrões genéricos e não informação específica de uma pessoa.

5.2.2 Seleção de um Conjunto de Dados Alvo

Como utilizou-se somente um Banco de Dados Relacional, a seleção do conjunto de dados alvo torna-se uma seleção de tabelas. Como o objetivo final da pesquisa é encontrar padrões sobre doenças, selecionaram-se algumas tabelas que tinham correlação com os beneficiários, tais como procedimentos médicos e diagnósticos, para extrair dados que possam satisfazer esse objetivo. Analisando essa base de dados e levando em consideração o objetivo de encontrar padrões sobre doenças, selecionaram-se as seguintes tabelas, GRS-BENEFICIARIO, GRS-GUIA, GRS-GUIA-EVENTO, SAM-TGE e SAM-CID cujas descrições seguem abaixo.

- **GRS-BENEFICIARIO** - Todos os beneficiários inscritos no plano de saúde.
- **GRS-GUIA** - Possui dados sobre os processos administrativos e de diagnósticos dos beneficiários.
- **GRS-GUIA-EVENTO** - Possui eventos utilizados pelos beneficiários.

- **SAM-TGE** - É a tabela geral de eventos e possui a descrição de todos os eventos possíveis do plano.
- **SAM-CID** - Essa tabela contém todos os códigos possíveis de CID (Classificação Internacional de Doenças), além de sua descrição.

5.2.3 Limpeza dos dados

No mundo real os dados geralmente são incompletos, inconsistentes e apresentam algum tipo de ruído. Limpeza de dados visa detectar e remover essas anomalias dos dados com o objetivo de aumentar/melhorar sua qualidade (Neves, 2003).

Na etapa de seleção do conjunto de dados alvo foi criada uma nova tabela com o histórico dos beneficiários na qual já foi realizada alguma limpeza nos dados, em que se levou em consideração, por exemplo, se um beneficiário usou ou não uma especialidade ou teve ou não uma determinada doença. Isso eliminou problemas com valores nulos em muitos atributos, pois considerou-se, por exemplo, se um paciente teve ou não determinada doença.

Segundo Neves (2003), para atingir o objetivo de descobrir conhecimento útil a partir de grandes bases de dados é importante considerar a possibilidade de retornar e avançar entre as fases do processo de DCBD repetindo certas etapas para obter um melhor resultado. Portanto, no decorrer do pré-processamento foram detectadas anomalias e ruídos nos dados como, a detecção de idades com valores negativos, implicando que a etapa de limpeza dos dados fosse aplicada novamente.

5.2.4 Pré-seleção de atributos

Para se realizar uma seleção prévia dos atributos foi levado em consideração o fato de muitos atributos relevantes praticamente não possuírem dados, como o caso do atributo IMC (Índice de Massa Corpórea) da tabela GRS-BENEFICIÁRIO. Após essa pré-seleção de atributos foram realizadas entrevistas com analistas que desenvolvem sistemas para as operadoras de planos de saúde, visando esclarecer dúvidas em relação aos atributos pré-selecionados, e após essas entrevistas foram descartados mais alguns atributos. As tabelas 5.1, 5.2, 5.3, 5.4 e 5.5 apresentam os atributos selecionados.

Com base nas tabelas 5.1 a 5.5 foi realizada a junção entre as mesmas e conseguiu-se retirar um histórico de procedimentos dos pacientes. Para fazer a junção encontrou-se vários problemas, como por exemplo, cada beneficiário pode possuir várias guias e cada guia pode conter vários eventos. Isso faria, após o processo de junção, um beneficiário

Tabela 5.1: Tabela de Beneficiário, GRS-BENEFICIARIO

Atributo	**Tipo**
HANDLE	INTEGER
SEXO	CHAR(1)
DATANASCIMENTO	DATETIME
ESTADO	INTEGER
MUNICÍPIO	INTEGER
SITUAÇÃO	INTEGER
TIPO	INTEGER

Tabela 5.2: Tabela de Guia, GRS-GUIA

Atributo	**Tipo**
HANDLE	INTEGER
BENEFICIARIO	INTEGER
GUIA	VARCHAR(20)
TIPOGUIA	INTEGER
CID01	INTEGER
ESPECIALIDADEDESCRICAO	INTEGER

Tabela 5.3: Tabela de Guia Evento, GRS-GUIA-EVENTO

Atributo	**Tipo**
HANDLE	INTEGER
BENEFICIARIO	INTEGER
GUIA	VARCHAR(20)
EVENTO	INTEGER

Tabela 5.4: Tabela Geral de Eventos, SAM-TGE

Atributo	**Tipo**
HANDLE	INTEGER
ESTRUTURA	VARCHAR(20)
DESCRICAO	VARCHAR(20)

Tabela 5.5: Tabela de códigos de CID, SAM-CID

Atributo	**Tipo**
HANDLE	INTEGER
DESCRICAO	VARCHAR(200)
ESTRUTURA	VARCHAR(5)

possuir vários registros o que impossibilitaria a ideia de histórico, ou seja, haveriam vários registros de beneficiários para cada guia, e vários registros de guias para cada evento o que invalidaria a etapa de mineração de dados com muitos registros repetidos.

Com base na junção preliminar das tabelas GRS-BENEFICIARIO, GRS-GUIA e GRS-GUIA-EVENTO, obteve-se uma tabela maior contendo os seguintes atributos.

- **BENEFICIARIO** - Código identificador do beneficiário.
- **SEXO** - Sexo do beneficiário.
- **IDADE** - Idade do beneficiário
- **ESTADO** - Estado onde o beneficiário reside.
- **MUNICIPIO** - Município onde o beneficiário reside.
- **TIPO-BENEF** - Se o beneficiário é titular ou dependente.
- **TIPO-GUIA** - Se o beneficiário fez uma consulta, usou medicamentos, teve uma internação ou se utilizou do SADT (Serviço Auxiliar de Diagnóstico e Terapia).
- **ESPECIALIDADE** - Indica a especialidade utilizada pelo beneficiário para uma determinada GUIA.
- **EVENTO** - Indica o evento que o beneficiário utilizou em uma determinada GUIA.
- **CID** - Classificação Internacional de Doenças - Identifica qual doença o beneficiário foi diagnosticado.

Como cada beneficiário pode utilizar muitas guias, várias especialidades, muitas CIDs e muitos eventos, fez-se necessário transformar os atributos TIPO-GUIA, ESPECIALIDADE, EVENTO e CID em vários outros atributos.

O atributo TIPO-GUIA possui como domínio quatro valores, SADT, CONSULTA, INTERNAÇÃO e MEDICAMENTOS. O valor MEDICAMENTOS não possui nenhum registro na base de dados, sendo portanto descartado. A partir do atributo TIPO-GUIA obteve-se os atributos SADT, CONSULTA e INTERNAÇÃO no qual os valores possíveis é a quantidade de utilização de cada serviço.

O atributo ESPECIALIDADE possui como domínio 69 especialidades diferentes. Então a partir desse atributo obteve-se 69 novos atributos que correspondem a cada especialidade disponibilizada pelo plano de saúde. O domínio de cada novo atributo contém a quantidade de vezes que um beneficiário utilizou a especialidade.

O atributo EVENTO possui como domínio 17.324 eventos o que torna difícil a transformação desses valores em atributos. Entretanto as tabelas GRS-GRUPOEVENTOS e GRS-GRUPOEVENTOS-EVENTO agrupam eventos que se relacionam a alguma doença

comum como, Diabetes Mellitus ou Hipertensão Arterial. A partir desses grupos criou-se então 11 atributos e, de acordo com os eventos realizados pelos beneficiários, incrementava-se o valor para o grupo no qual o evento pertencia. O domínio desses grupos é o número de vezes que cada beneficiário utilizou os eventos pertencentes a um determinado grupo de eventos.

O atributo CID possui como domínio 14.202 códigos CID, fato esse que torna difícil a transformação desses valores em atributos. Entretanto, encontrou-se na literatura (de Castro e Carvalho, 2005) um agrupamento dessas CIDs, o qual foi construído de forma empírica tendo como meta principal manter a coerência clínica dentro do mesmo capítulo da CID e um volume mínimo de internações em cada novo grupo para análise de re-internações hospitalares. Segundo os autores, apesar de ter foco em internação ou re-internações hospitalares, esses agrupamentos podem ser úteis para análise de problemas de saúde e também podem ser adaptados para outros objetivos específicos de pesquisa. Utilizou-se esses agrupamentos para transformação desses dados em atributos totalizando 32 novos atributos, no qual o valor de cada atributo é o número de vezes que o paciente foi atendido com um determinado código de CID pertencente a um grupo de CID.

No Apêndice A apresenta-se a descrição do significado de cada atributo gerado pela pré-seleção de atributos.

5.2.5 Redução e projeção de dados

Para se ter uma boa redução dos dados faz-se necessário conhecer como esses dados estão organizados, qual relevância de cada atributo e também qual a relação entre os atributos. Para ter esse conhecimento dos dados foi necessário fazer algumas análises estatísticas. Nas próximas seções são apresentadas essas análises.

Classificação Internacional de Doenças (CID)

Devido ao fato do atributo CID ser convertido em vários atributos, apresentam-se alguns gráficos para melhor entendimento desses novos atributos gerados. A Figura 5.1 demonstra um comportamento desses novos atributos.

A partir do gráfico da Figura 5.1 percebeu-se que o agrupamento XXI (Contato com os serviços de saúde) concentra a maior parte do total de códigos, 65,72% das ocorrências ante 34,28% dos demais grupos. Esse grupo é formado pelos códigos que pertencem ao intervalo Z00 a Z99. Em uma consulta ao banco de dados verificou-se que mais de 98% desses registros pertencem a CID Z00 que segundo o *website* do Ministério da Saúde

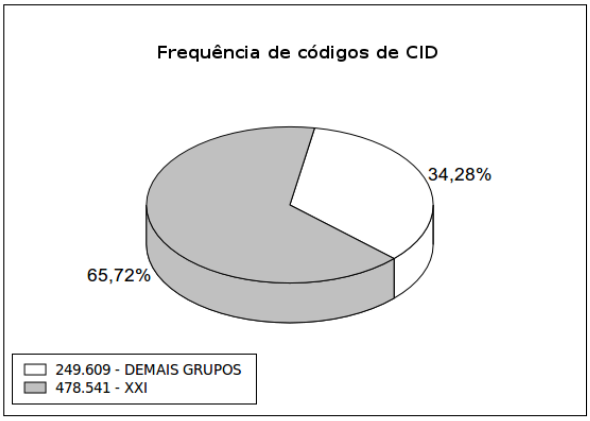

Figura 5.1: Frequência do grupo CID XXI em relação aos demais grupos

(MS, 2011) correspondem a exames em geral e investigações de pessoas sem queixas ou diagnósticos relatados.

Em entrevistas com os analistas de domínio constatou-se que a maior parte desses códigos são casos nos quais os médicos não informam o CID do paciente ou não houve diagnóstico.

Pela alta proporção do atributo GRUPO-CID-XXI em relação aos demais atributos retirou-se o grupo XXI da geração do histograma de frequência de grupos de códigos de CID para uma melhor visualização e comparação.

Verifica-se, a partir do histograma da Figura 5.2, que há vários grupos de códigos de CID com baixa frequência o que pode dificultar ou inviabilizar o processo de descoberta de padrões levando em consideração esses atributos.

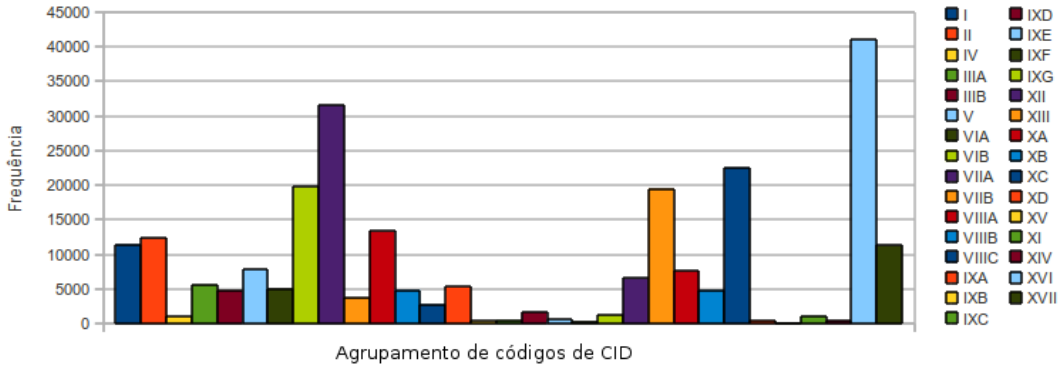

Figura 5.2: Histograma de frequência de agrupamentos de códigos de CID

Como o processo de extração de dados procurou preservar informações como a quantidade de vezes que um beneficiário foi internado ou a quantidade de vezes que

foi tratar-se de uma determinada doença “CID”, fez-se uma nova análise dos dados agora desconsiderando o número de vezes que o beneficiário foi atendido com uma determinada doença, considerando apenas se teve uma determinada doença ou não. O gráfico da Figura 5.3 mostra que a proporção entre os demais grupos de códigos e o grupo XXI se inverte.

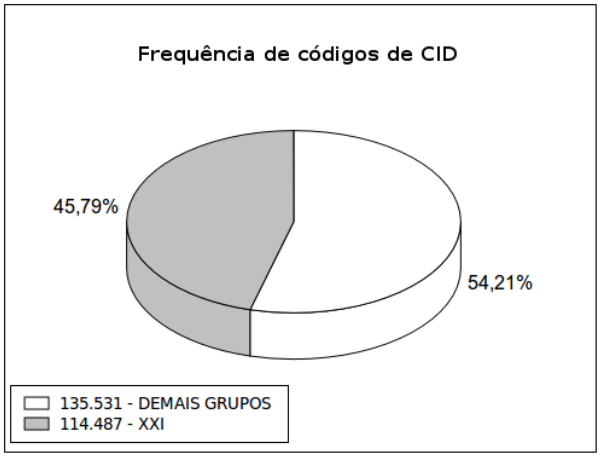

Figura 5.3: Frequência do grupo CID XXI em relação aos demais grupos desconsiderando a quantidade de atendimentos

Mesmo havendo uma redução da proporção em relação ao total de registros o grupo XXI ainda possui uma alta proporção, o que atrapalha a visualização de um histograma que compara a frequência de cada grupo de CID.

O histograma da Figura 5.4 leva em consideração se um beneficiário teve ou não uma determinada doença. Percebe-se que o comportamento do histograma da Figura 5.4 é muito similar ao histograma da Figura 5.2. Isso significa que os grupos com baixa frequência continuam com baixa frequência em relação aos de alta frequência o que pode ser um indicativo de quais grupos seriam melhores para a aplicação dos algoritmos de MD.

Especialidades

Como o atributo ESPECIALIDADE foi convertido em 69 atributos, fez-se necessário uma análise para melhor entendimento dos novos atributos gerados. Entretanto, nessa análise não foi possível criar histogramas de frequência comparando todas as especialidades, pois há muitas especialidades. O gráfico da Figura 5.5 destaca um dos comportamentos dos atributos relativos a especialidade.

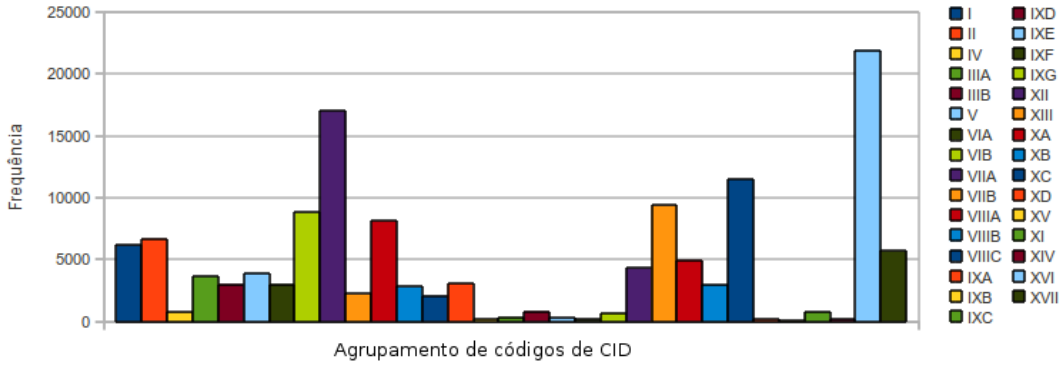

Figura 5.4: Histograma de frequência de agrupamentos de códigos de CID desconsiderando repetições

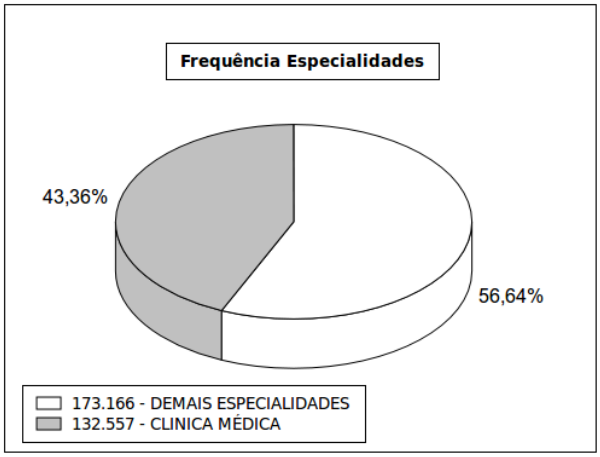

Figura 5.5: Frequência de Clínica Médica em relação as demais especialidades

O gráfico da Figura 5.5 mostra um caso similar ao ocorrido nos grupos de códigos de CID, no qual um atributo possui uma proporção extremamente maior que os demais atributos. Pela alta proporção do atributo CLÍNICA MÉDICA é conveniente descartá-lo para construção de um histograma de frequência com todas as especialidades. Entretanto, devido ao alto número de atributos, a construção de histogramas tornou-se inviável e fez-se necessário exibir as frequências dos atributos por meio de uma tabela.

No Apêndice B encontra-se a tabela de frequência dos atributos de especialidade, na qual percebe-se que vários atributos de especialidade possui baixa cobertura, o que pode dificultar a etapa mineração de dados, pois aumenta a dificuldade em encontrados padrões de alta qualidade e cobertura satisfatória considerando esses atributos (Monard e Baranauskas, 2003). Também são exibidas as frequências desconsiderando repetições,

o que em alguns casos mostra uma grande variação, como pode ser vista no atributo ESPECIALIDADE-18 e pelo gráfico da Figura 5.6.

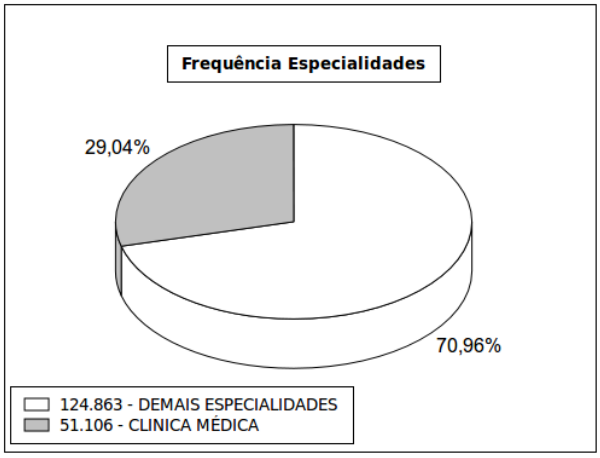

Figura 5.6: Frequência de Clínica Médica em relação as demais especialidades desconsiderando repetições

SADT, Consultas e Internações

Segundo a Agência Nacional de Saúde (ANS, 2011), a Guia de Serviços Profissionais / Serviço Auxiliar Diagnóstico e Terapia (SP/SADT) deve ser utilizada no atendimento a diversos tipos de eventos, tais como: remoção, pequena cirurgia, terapias, consulta com procedimentos, exames, atendimento domiciliar, SADT internado ou quimioterapia, radioterapia ou terapia renal substitutiva (TRS). A Guia de Consultas deve ser utilizada exclusivamente na execução de consultas eletivas sem procedimentos, já a Guia de Solicitação de Internação é referente aos casos de internação. Essas informações estão presentes nos dados por meio dos atributos SADT, CONSULTAS e INTERNACAO, que estão representadas na Figura 5.7.

Fica claro no histograma da Figura 5.7 que o número de consultas é o maior entre os três procedimentos, e o número de internações é bem mais baixo que o número de consultas e SADT. Esse histograma considera o número de vezes que cada usuário utilizou cada serviço.

O histograma da Figura 5.8 demonstra a relação entre esses três atributos sem considerar repetições. Nesse histograma é possível observar que a proporção de internações e SADT diminuem consideravelmente em relação a consultas. Pode-se observar também que há um alto número de reinternações, pois no gráfico da Figura 5.8 observa-se que

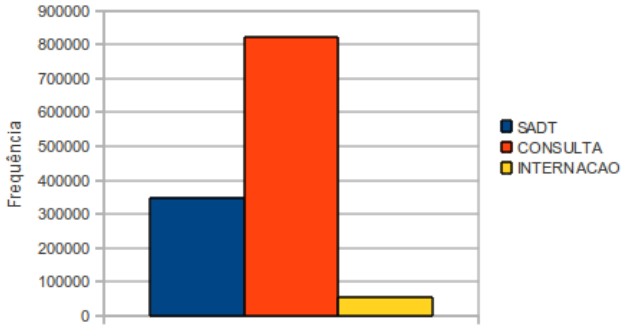

Figura 5.7: Frequência de SADT, Consulta e Internação

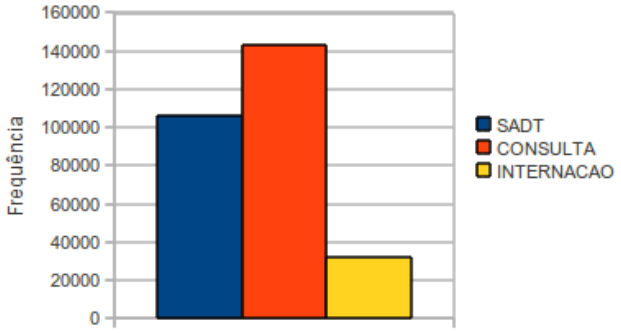

Figura 5.8: Frequência de SADT, Consulta e Internação sem repetições

31.617 beneficiários já sofreram internação e no gráfico da Figura 5.7 que há registrado 56.560 internações, o que resulta em 24.943 reinternações.

Sexo, Idade, Cidade e Estado

A Figura 5.9 demonstra a proporção de beneficiários do sexo masculino e feminino, no qual é possível ver que a proporção de mulheres é relativamente maior que a de homens.

A idade dos beneficiários varia entre -1 e 105 anos com uma concentração maior em torno dos 50 anos. A idade -1 corresponde a um ruído nos dados e podem ser casos de idade não informada ou algum problema no processo de entrada de dados e corresponde a 0,38% do total de registros. Nesses casos os registros foram desconsiderados.

A maioria dos beneficiários (85,74%) dos 159.883 pertence a um mesmo estado (Santa Catarina). Os demais estados possuem quantidade irrisória e 95% dos 22.794 beneficiários restantes não possuem informação de estados em seus cadastros. O atributo

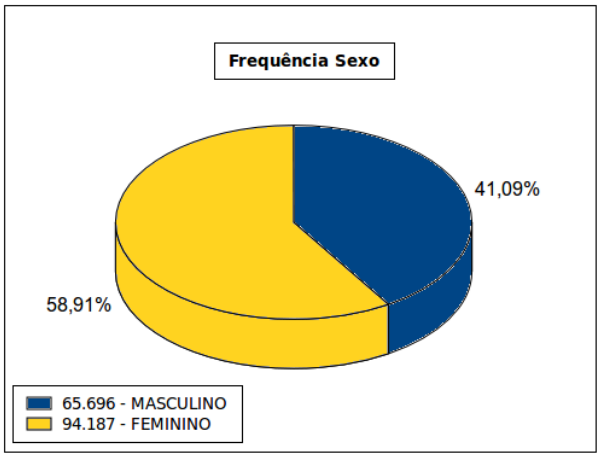

Figura 5.9: Frequência do atributo Sexo

MUNICÍPIO possui um domínio com 356 cidades, com maior concentração em uma região metropolitana, composta por duas cidades, com 32.301 beneficiários. Há também 14% do total de beneficiários que não possuem suas cidades informadas.

Grupo de eventos

Eventos são procedimentos realizados pelos beneficiários dentro ou fora dos hospitais, por exemplo, um medicamento que deve ser aplicado. O alto número de eventos cadastrados na base tornou difícil a análise desses dados na forma como eram armazenados. Estudando a base de dados percebeu-se que esses eventos sofriam um agrupamento, ou seja, buscava-se agrupar eventos que são realizados por beneficiários com algum problema de saúde. Por exemplo, o agrupamento "procedimentos relacionados a câncer de mama" procura determinar quais eventos estão relacionados as pessoas que tenham câncer de mama.

Os histogramas das Figuras 5.10 e 5.11 demonstram as frequências desses agrupamentos de eventos.

Percebe-se pelo histograma da Figura 5.10 que alguns grupos não possuem ou quase não possuem dados o que inviabiliza a utilização dos mesmos para MD.

Quando se desconsidera as repetições, como no histograma da Figura 5.11, a proporção de dados diminui, permitindo a seleção de alguns grupos de eventos.

Seleção de atributos relevantes

Segundo Han e Kamber (2006), a seleção de atributos em um conjunto de dados reduz o tamanho do conjunto removendo atributos irrelevantes ou redundantes. Embora seja

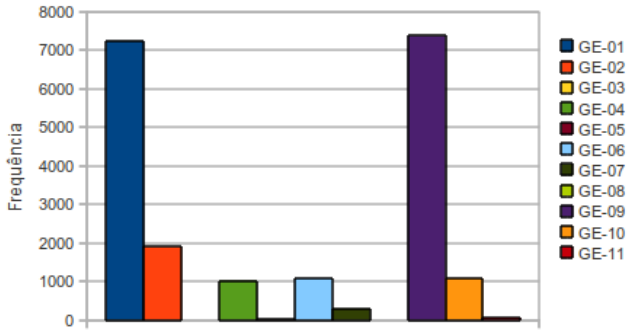

Figura 5.10: Frequência de Grupos de Eventos

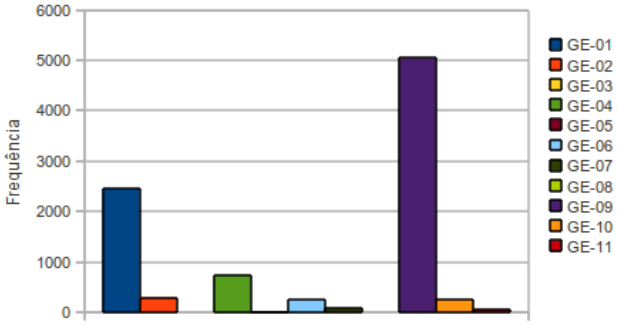

Figura 5.11: Frequência de Grupos de Eventos sem repetições

possível para um especialista de domínio escolher alguns atributos úteis, isso pode ser uma tarefa complicada e consumir muito tempo quando o comportamento dos dados não é conhecido. Deixar atributos relevantes fora do conjunto de atributos selecionados ou manter atributos irrelevantes pode causar confusão para o algoritmo de MD. Isso pode resultar na descoberta de padrões de baixa qualidade. Além disso, o alto número de atributos irrelevantes ou redundantes aumenta o custo computacional do processo de DCBD.

Um conceito importante considerado para seleção dos atributos foi o conceito de Prevalência de Classe (Monard e Baranauskas, 2003), referente ao desbalanceamento de classes. Por exemplo, considere um conjunto com a seguinte distribuição de classes dist(C1,C2,C3) = (99,00%, 0,25%, 0,75%), com prevalência da classe C1. Um classificador simples que classifique sempre novos exemplos como pertencentes à classe majoritária C1 teria uma precisão de 99,00%. Isso pode ser indesejável quando as classes minoritárias possuem alguma informação muito importante. Por exemplo, supondo C1: paciente

normal, C2: paciente com doença A e C3: paciente com doença B. Isso ocorre, por exemplo, no GRUPO-CID-VIIIC (Doenças crônicas das vias aéreas inferiores) que possui uma distribuição de classes dist(SIM, NÃO) = (1,30%, 98,70%), em que SIM: paciente com doença e NÃO: paciente normal, com prevalência da classe NÃO. No caso desse atributo, o número de registros da classe SIM é igual a 2.071, valor pouco significativo no universo de 159.883 registros desse conjunto de dados ou em relação aos demais agrupamentos de códigos de CID. Portanto, esse atributo foi eliminado do conjunto de atributos selecionados. Esse conceito foi aplicado na seleção de todos os atributos.

Com base nas análises das seções anteriores e o conceito de prevalência de classe fez-se a seleção dos atributos para MD. Primeiramente fez-se a seleção dos grupos de códigos de CID. Com base nos histogramas das Figuras 5.2 e 5.4, selecionou-se os atributos com maior frequência da classe SIM e descartou-se os atributos com frequência menor que 2500 registros da classe SIM, devido a diferença desproporcional relativa aos demais atributos, como visto pelo histograma da Figura 5.4. O mesmo procedimento foi efetuado para especialidade, mas foram eliminadas as especialidades com frequência menor que 700 ocorrências também devido a diferença desproporcional da classe SIM, essa diferença pode ser vista por meio da tabela do Apêndice B. Os atributos SADT, Consultas e Internações foram mantidas pela sua relevância e alto número de ocorrências, já os atributos Sexo, Idade, Cidade e Estado também foram mantidos por possuírem informações importantes em relação aos beneficiários. Em relação aos grupos de eventos selecionou-se os grupos 01 (Procedimentos relacionados a Diabetes Mellitus), 04 (Procedimentos Sentinelas relacionadas a Diabetes Meliltus) e 09 (Procedimentos relacionados ao Câncer de Mama) por possuírem alta frequência. A Tabela 5.6 apresenta os atributos selecionados.

Tabela 5.6: Descrição dos atributos selecionados

Atributo	**Domínio**	**Descrição**
SEXO	F, M	Sexo do beneficiário
IDADE	[0..105]	Idade do beneficiário
ESTADO	SC, RS, PR, SP, RN, MG, RJ, DF, AL, MT, GO	Estado do beneficiário
MUNICÍPIO	[ABDON BATISTA..ZORTEA]	Município do beneficiário
SADT	[0..285]	Serviço Auxiliar de Diagnóstico e Terapia

Tabela 5.6: Descrição dos atributos selecionados

Atributo	**Domínio**	**Descrição**
CONSULTA	[0..101]	Consulta médica
INTERNAÇÃO	[0..63]	Internação hospitalar
GRUPO-CID-I	[0..69]	Doenças infecciosas e parasitárias
GRUPO-CID-II	[0..95]	Neoplasias
GRUPO-CID-IIIA	[0..41]	Doenças endócrinas
GRUPO-CID-IIIB	[0..49]	Doenças nutricionais e metabólicas
GRUPO-CID-V	[0..272]	Doenças mentais
GRUPO-CID-VIA	[0..175]	Sistema nervoso
GRUPO-CID-VIB	[0..80]	Olhos e anexos, ouvido e apófise mastóide
GRUPO-CID-VIIA	[0..144]	Circulatório, menos veias e linfáticos
GRUPO-CID-VIIIA	[0..74]	Infecções respiratórias agudas
GRUPO-CID-VIIIB	[0..64]	Outras doenças respiratórias
GRUPO-CID-IXA	[0..138]	Doenças do esôfago, estômago e duodeno
GRUPO-CID-XII	[0..48]	Pele e tecido subcutâneo
GRUPO-CID-XIII	[0..115]	Osteomuscular e tecido conjuntivo
GRUPO-CID-XA	[0..56]	Doenças urinárias
GRUPO-CID-XB	[0..85]	Doenças genitais masculinas
GRUPO-CID-XC	[0..117]	Doenças genitais femininas
GRUPO-CID-XVI	[0..93]	Sintomas, sinais e afecções mal definidas
GRUPO-CID-XVII	[0..82]	Lesões, envenenamentos e causas externas
GRUPO-CID-XXI	[0..66]	Contato com os serviços de saúde
ESPECIALIDADE-04	[0..16]	Análises clínicas
ESPECIALIDADE-05	[0..8]	Anatomia patológica
ESPECIALIDADE-06	[0..6]	Anestesiologia
ESPECIALIDADE-07	[0..7]	Angiologia
ESPECIALIDADE-09	[0..16]	Cardiologia
ESPECIALIDADE-13	[0..8]	Cirurgia geral
ESPECIALIDADE-18	[0..56]	Clínica médica

Tabela 5.6: Descrição dos atributos selecionados

Atributo	**Domínio**	**Descrição**
ESPECIALIDADE-24	[0..7]	Dermatologia
ESPECIALIDADE-27	[0..4]	Endocrinologia
ESPECIALIDADE-31	[0..5]	Gastroenterologia
ESPECIALIDADE-33	[0..12]	Ginecologia
ESPECIALIDADE-39	[0..23]	Internações
ESPECIALIDADE-47	[0..12]	Medicina preventiva
ESPECIALIDADE-50	[0..5]	Neurologia
ESPECIALIDADE-53	[0..9]	Oftalmologia
ESPECIALIDADE-54	[0..11]	Ortopedia
ESPECIALIDADE-55	[0..9]	Otorrinolaringologia
ESPECIALIDADE-56	[0..6]	Patologia
ESPECIALIDADE-57	[0..11]	Patologia clínica
ESPECIALIDADE-58	[0..13]	Pediatria
ESPECIALIDADE-59	[0..6]	Pneumologia
ESPECIALIDADE-62	[0..6]	Psiquiatria
ESPECIALIDADE-63	[0..5]	Radiodiagnostico
ESPECIALIDADE-64	[0..9]	Radiologia
ESPECIALIDADE-66	[0..4]	Reumatologia
ESPECIALIDADE-69	[0..9]	Urologia
GRUPO-EVENTO-01	[0..178]	Procedimentos - Diabetes mellitus
GRUPO-EVENTO-04	[0..19]	Procedimentos sentinelas - Diabetes mellitus
GRUPO-EVENTO-09	[0..45]	Procedimentos câncer de mama

5.3 Considerações Finais

Com o objetivo de obter dados preparados para MD, aplicou-se a metodologia descrita neste capítulo para selecionar os principais atributos, além de eliminar os registros inúteis ou com problemas nos dados. Com a aplicação dessa metodologia obteve-se uma redução de 120 para 55 atributos e redução em 14% no número de registros de um total de 159.883 registros considerados. A maneira como os dados foram re-organizados facilita a aplicação

de algoritmos de MD e ainda podem sofrer algum processo de refinamento dependendo do foco que se queira dar a utilização desses dados.

Houve algumas dificuldades para compreensão do domínio da aplicação, as quais foram superadas por meio da análise da estrutura de dados e estudo da documentação da estrutura, obtendo-se informações como o significado de tabelas e atributos além de conhecimentos referentes aos processos de saúde suplementar. Utilizou-se também análise estatística baseada no uso de tabelas e gráficos.

Com os estudos feitos sobre os dados foi possível obter uma visão mais clara do domínio da aplicação. Essa visão facilitou a seleção de atributos e o entendimento de alguns comportamentos como, por exemplo, casos em que pacientes que sofrem internação estão mais propensos a sofrerem novas internações. Isso foi constatado por meio dos gráficos de frequência.

A utilização de gráficos facilitou a escolha dos atributos e ajudou na eliminação de atributos de baixa frequência, que provocaria a descoberta de regras de baixa qualidade devido a baixa cobertura desses atributos, o que ocasionaria um grande esforço computacional desnecessário na aplicação dos algoritmos de MD pois aumentaria o espaço de busca.

Capítulo

6

Experimentos e Resultados

6.1 Considerações Iniciais

A aplicação de algoritmos de mineração de dados exige, além dos próprios dados, a interação do analista/programador com especialistas de domínio para conhecer o conteúdo, significado e características dos dados disponíveis. Com utilização de métodos *user-driven*, a interação com o especialista de domínio deve ser ainda maior principalmente no que diz respeito as impressões gerais utilizadas nesta dissertação.

Neste capítulo são apresentados experimentos para avaliação do algoritmo de programação genética e do algoritmo cultural em que são utilizadas impressões gerais como conhecimento de domínio.

6.2 Descrição Geral dos Experimentos

O conjunto de dados utilizado foi previamente preparado, conforme descrito no Capítulo 5. Esse conjunto, contendo informações relativas a procedimentos realizados por beneficiários de um plano de saúde, possui 137.020 registros e 55 atributos e seguem o padrão *arff*, o mesmo padrão utilizado pela ferramenta Weka (Witten e Frank, 2005). Todos os grupos de códigos de CID e especialidades foram categorizadas como "Sim" para uma ou mais ocorrências e "Não" para nenhuma ocorrência. O atributo idade foi discretizado em intervalos de 10 anos e o atributo internação foi categorizado em "nenhuma" para nenhuma internação, "uma" para apenas uma internação e "mais_que_uma" para mais que uma

internação. Os demais atributos permaneceram com o mesmo domínio apresentado na Tabela 5.6. Os únicos atributos que mantiveram o tipo numérico são SADT e consulta, e tiveram seus parâmetros difusos definidos como $p1 = 2$, $p2 = 4$, $p3 = 6$ e $p4 = 8$ para SADT e $p1 = 0$, $p2 = 1$, $p3 = 3$ e $p4 = 5$ para consulta.

Os atributos selecionados para os experimentos são apresentados pela Tabela 6.1 e a distribuição de suas classes pela Tabela 6.2, e foram selecionados de acordo com o interesse dos especialistas de domínio. A distribuição das demais classes podem ser encontradas no Apêndice C.

Tabela 6.1: Atributos selecionados

Atributo	**Classe**	**Descrição**
GRUPO-CID-IIIA	Sim, Não	Doenças Endócrinas
GRUPO-CID-IIIB	Sim, Não	Doenças Nutricionais e Metabólicas
GRUPO-EVENTO-09	Sim, Não	Procedimentos Câncer de Mama

Tabela 6.2: Distribuição de classes

Atributo	**Descrição**	**SIM(%)**	**NÃO(%)**
GRUPO-CID-IIIA	Doenças Endócrinas	2,26	97,74
GRUPO-CID-IIIB	Doenças Nutricionais e Metabólicas	1,85	98,16
GRUPO-EVENTO-09	Procedimentos Câncer de Mama	3,62	96,39

Foram realizados experimentos sobre o algoritmo de programação genética com o objetivo de validar o mesmo. Também, foram realizados experimentos com o algoritmo cultural no qual as impressões gerais são utilizadas de duas maneiras diferentes.

6.3 Primeiro Experimento

O primeiro experimento se refere a aplicação do algoritmo de PG, apresentado no Capítulo 3, sobre os dados da saúde. Para realizar esse experimento foi necessário calibrar os parâmetros do algoritmo e para isso utilizou-se a metologia apresentada por de Pinho *et al.* (2007) que utiliza um projeto experimental fatorial completo para calibragem de parâmetros de um algoritmo genético.

A metodologia apresentada por de Pinho *et al.* (2007) é composta de sete passos e a condução desses passos são apresentados a seguir.

6.3.1 Definição do Problema

O problema é otimizar a função de aptidão utilizada pelo algoritmo de PG. A função de aptidão é apresentada pela Equação 3.20. O objetivo deste projeto de experimentos é identificar qual a escolha dos parâmetros do algoritmo de PG dará a melhor qualidade à resposta.

6.3.2 Escolha de Fatores, Níveis e Variações

Os fatores selecionados para o projeto de experimentos são os parâmetros do algoritmo de PG (tamanho da população, número de gerações, taxa de cruzamento e taxa de mutação). Optou-se em trabalhar com dois níveis para cada fator. Esses níveis e suas variações são apresentados na Tabela 6.3.

Tabela 6.3: Fatores, Níveis e Variações do Projeto de Experimentos

		Níveis	
Fatores		-	+
A	Tamanho da População	50	100
B	Número de Gerações	50	100
C	Taxa de Cruzamento	0,6	0,9
D	Taxa de Mutação	0,01	0,3

6.3.3 Seleção da Variável de Resposta

A variável de resposta para esse projeto de experimentos será o valor maximizado da aptidão do indivíduo apresentado anteriormente.

6.3.4 Escolha do Projeto Experimental

Este projeto possui quatro fatores com dois níveis cada. Dessa forma, optou-se pelo uso do fatorial completo pois, de acordo com de Pinho *et al.* (2007), o uso de um fatorial completo dá maior eficiência em estimar os efeitos principais dos fatores sob a variável em análise e, principalmente, na definição da interação entre todos os fatores.

A matriz com o planejamento dos fatores é exibida pela Tabela 6.4. Para um experimento de fatorial completo com quatro fatores com dois níveis cada são necessários dezesseis experimentos ($2^4 = 16$). As colunas A, B, C e D presentes na Tabela 6.4 são referentes aos fatores descritos na Subseção 6.3.2.

Tabela 6.4: Matriz com o planejamento dos experimentos

Experimentos	A	B	C	D
1	-	-	-	-
2	+	-	-	-
3	-	+	-	-
4	+	+	-	-
5	-	-	+	-
6	+	-	+	-
7	-	+	+	-
8	+	+	+	-
9	-	-	-	+
10	+	-	-	+
11	-	+	-	+
12	+	+	-	+
13	-	-	+	+
14	+	-	+	+
15	-	+	+	+
16	+	+	+	+

6.3.5 Execução do Experimento

A execução do experimento foi realizado com base no algoritmo de PG apresentado no Capítulo 3, que foi implementado utilizando a linguagem *Java* no ambiente integrado de desenvolvimento *Eclipse 3.6* e aplicado sobre os dados apresentados anteriormente utilizando cada uma das classes apresentadas pela Tabela 6.1.

Para a execução do algoritmo os parâmetros do mesmo foram mantidos constantes durante toda a execução, sofrendo alteração somente após o critério de término do algoritmo ser alcançado seguindo a Tabela 6.4.

Os resultados encontrados em cada execução são apresentados pelas Tabelas 6.5, 6.6, e 6.7. Nota-se que para cada meta o algoritmo foi executado três vezes. Os resultados de cada uma das execuções são apresentadas pelas colunas Aptidão 1, Aptidão 2 e Aptidão 3 respectivamente.

6.3.6 Análise Estatística dos Dados

A partir dos resultados encontrados, calculou-se os efeitos principais de cada fator e o efeito de suas interações. Para os cálculos desenvolvidos utilizou-se o software MiniTab® versão 15.0.

Tabela 6.5: Execução do Experimento para a Classe "GRUPO-CID-IIIA = SIM"

Experimento	A	B	C	D	Aptidão 1	Aptidão 2	Aptidão 3	Média
1	-	-	-	-	0.072	0.065	0.063	0.066
2	+	-	-	-	0.234	0.234	0.378	**0.282**
3	-	+	-	-	0.147	0.071	0.1	0.106
4	+	+	-	-	0.149	0.225	0.352	0.242
5	-	-	+	-	0.073	0.119	0.134	0.108
6	+	-	+	-	0.167	0.178	0.114	0.153
7	-	+	+	-	0.159	0.12	0.176	0.151
8	+	+	+	-	0.181	0.132	0.078	0.130
9	-	-	-	+	0.228	0.228	0.174	0.210
10	+	-	-	+	0.177	0.222	0.318	0.239
11	-	+	-	+	0.147	0.166	0.231	0.181
12	+	+	-	+	0.419	0.241	0.185	0.281
13	-	-	+	+	0.153	0.07	0.176	0.133
14	+	-	+	+	0.234	0.222	0.133	0.196
15	-	+	+	+	0.182	0.125	0.241	0.182
16	+	+	+	+	0.307	0.147	0.24	0.231

Tabela 6.6: Execução do Experimento para a Classe "GRUPO-CID-IIIB = SIM"

Experimento	A	B	C	D	Aptidão 1	Aptidão 2	Aptidão 3	Média
1	-	-	-	-	0.047	0.039	0.057	0.047
2	+	-	-	-	0.056	0.062	0.047	0.055
3	-	+	-	-	0.044	0.065	0.042	0.050
4	+	+	-	-	0.058	0.05	0.066	0.058
5	-	-	+	-	0.052	0.073	0.057	0.060
6	+	-	+	-	0.053	0.066	0.048	0.055
7	-	+	+	-	0.043	0.042	0.078	0.054
8	+	+	+	-	0.12	0.114	0.064	0.099
9	-	-	-	+	0.066	0.054	0.066	0.062
10	+	-	-	+	0.083	0.188	0.1	0.123
11	-	+	-	+	0.049	0.054	0.052	0.051
12	+	+	-	+	0.119	0.067	0.248	0.144
13	-	-	+	+	0.054	0.19	0.063	0.102
14	+	-	+	+	0.318	0.054	0.067	**0.146**
15	-	+	+	+	0.059	0.248	0.133	**0.146**
16	+	+	+	+	0.06	0.064	0.054	0.059

Análise para o GRUPO-CID-IIIA

O peso dos efeitos para a classe "GRUPO-CID-IIIA = SIM" pode ser visualizado por meio da Figura 6.1. Nessa figura verifica-se que os fatores A, C e D assim como a interação entre os fatores AC e ACD são significativos para um nível de significância α de 5%.

Tabela 6.7: Execução do Experimento para a Classe "GRUPO-EVENTO-09 = SIM"

Experimento	A	B	C	D	Aptidão 1	Aptidão 2	Aptidão 3	Média
1	-	-	-	-	0.327	0.172	0.366	0.288
2	+	-	-	-	0.409	0.311	0.349	0.356
3	-	+	-	-	0.385	0.264	0.319	0.322
4	+	+	-	-	0.399	0.295	0.667	0.453
5	-	-	+	-	0.349	0.306	0.469	0.374
6	+	-	+	-	0.74	0.378	0.382	0.500
7	-	+	+	-	0.444	0.368	0.376	0.396
8	+	+	+	-	0.403	0.382	0.408	0.397
9	-	-	-	+	0.562	0.653	0.333	0.516
10	+	-	-	+	0.434	0.429	0.408	0.423
11	-	+	-	+	0.421	0.403	0.481	0.435
12	+	+	-	+	0.392	0.421	0.438	0.417
13	-	-	+	+	0.467	0.403	0.667	0.512
14	+	-	+	+	0.457	0.442	0.677	**0.525**
15	-	+	+	+	0.39	0.424	0.409	0.407
16	+	+	+	+	0.409	0.481	0.481	0.457

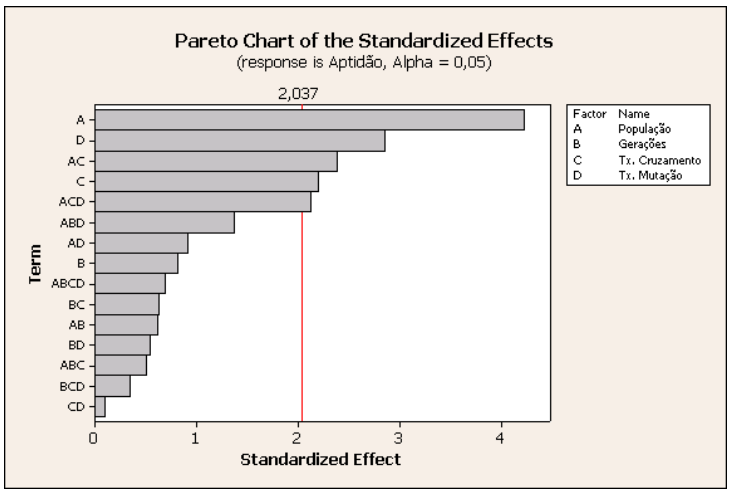

Figura 6.1: Gráfico de Pareto para os efeitos padronizados do GRUPO-CID-IIIA

A análise dos efeitos principais de cada fator, apresentada na Figura 6.2, mostra que os fatores "tamanho da população"e "taxa de mutação"possuem forte efeito positivo sobre a resposta final, ou seja, o valor máximo da função. Já o "número de gerações"possui baixa influência. O fator "taxa de cruzamento"possui efeito negativo sobre o resultado final.

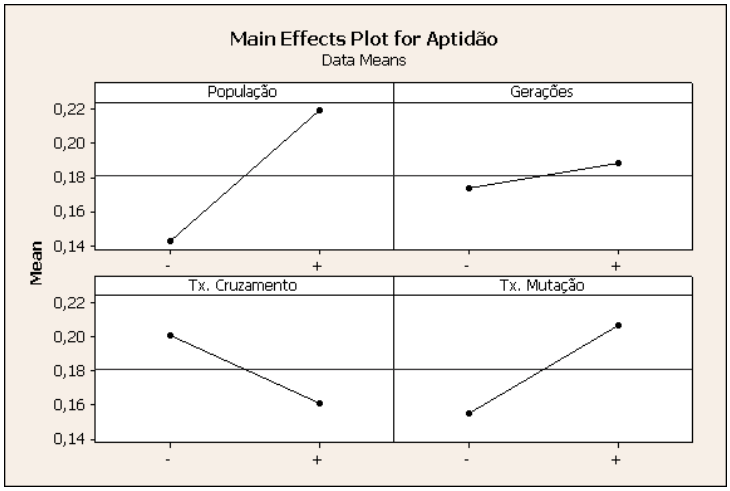

Figura 6.2: Gráfico dos efeitos principais para o GRUPO-CID-IIIA

Análise para o GRUPO-CID-IIIB

O peso dos efeitos para a classe "GRUPO-CID-IIIB = SIM" pode ser visualizado por meio da Figura 6.3. Nessa figura verifica-se que o fator D (taxa de mutação) é o único significativo para um nível de significância α de 5%.

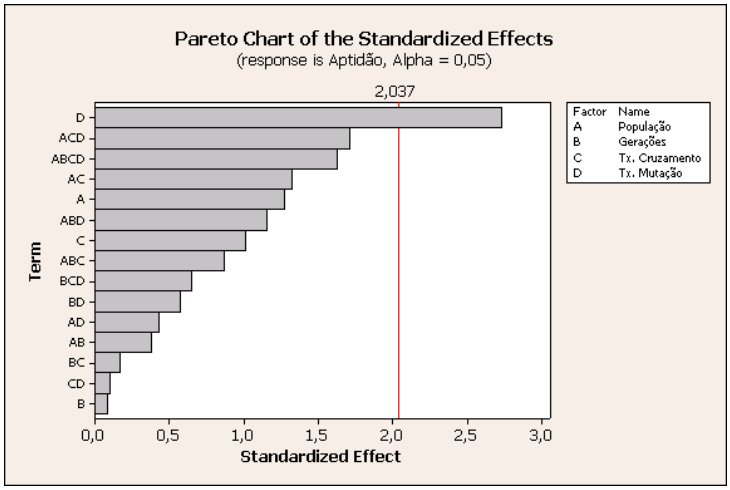

Figura 6.3: Gráfico de Pareto para os efeitos padronizados do GRUPO-CID-IIIB

A análise dos efeitos principais de cada fator, apresentada na Figura 6.4, mostra que o fator "taxa de mutação"possui forte efeito positivo sobre a resposta final, ou seja, o valor máximo da função. Já os fatores "taxa de cruzamento"e "tamanho da população"possuem influência não muito forte. O fator "número de gerações"não possui efeito significativo sobre o resultado final.

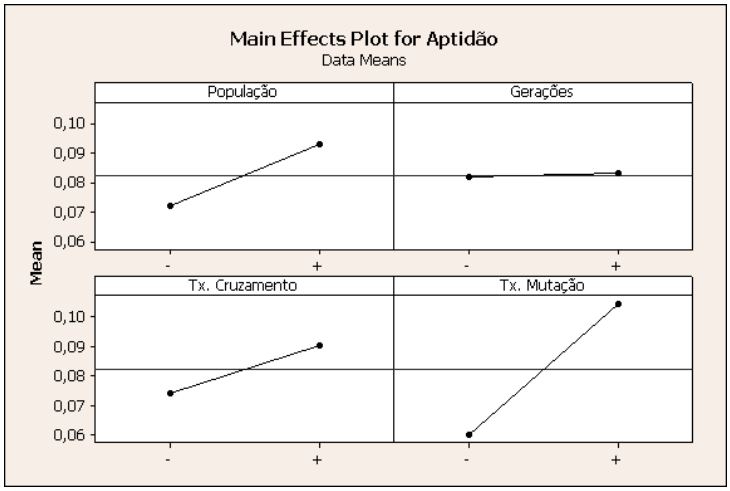

Figura 6.4: Gráfico dos efeitos principais para o GRUPO-CID-IIIB

Análise para o GRUPO-EVENTO-09

O peso dos efeitos para a classe "GRUPO-EVENTO-09 = SIM" pode ser visualizado por meio da Figura 6.5. Nessa figura verifica-se que o fator D (taxa de mutação), é o único significativo para um nível de significância α de 5%.

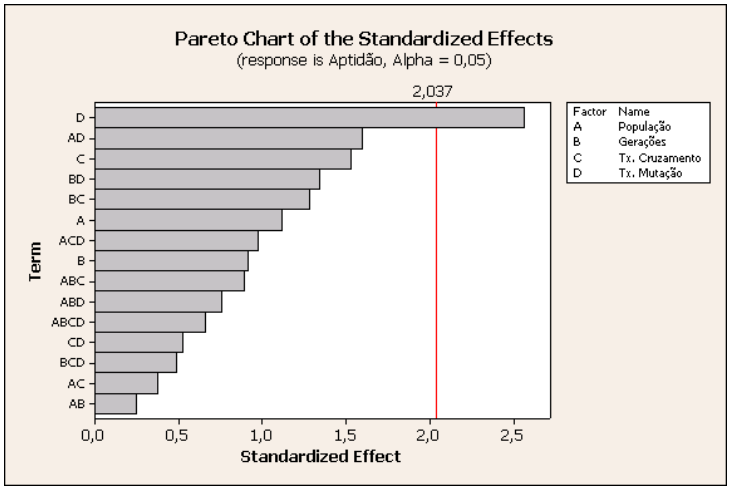

Figura 6.5: Gráfico de Pareto para os efeitos padronizados do GRUPO-EVENTO-09

A análise dos efeitos principais de cada fator, apresentada na Figura 6.6, mostra que o fator "taxa de mutação" possui forte efeito positivo sobre a resposta final, ou seja, o valor máximo da função. Já os fatores "tamanho da população"e "taxa de cruzamento"possuem influência não muito forte. O fator "número de gerações"possui efeito negativo moderado sobre o resultado final.

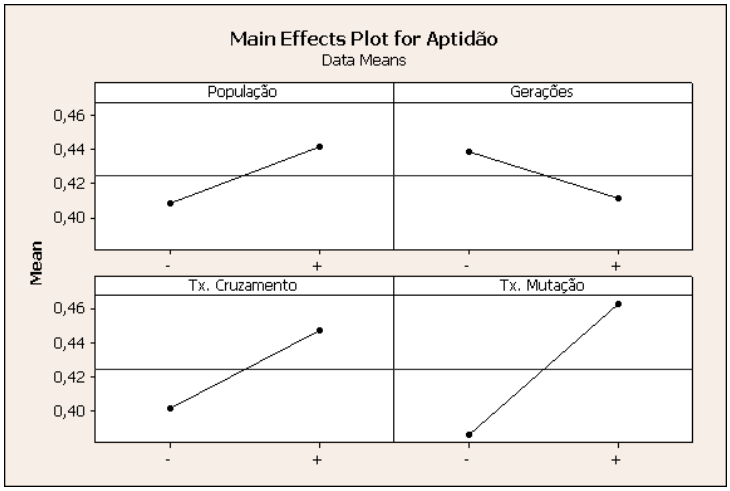

Figura 6.6: Gráfico dos efeitos principais para GRUPO-EVENTO-09

6.3.7 Conclusões e Recomendações

Com base nos resultados obtidos nos experimentos anteriores definiu-se os valores dos parâmetros do algoritmo para cada classe, conforme a Tabela 6.8, pois são os valores que mais se ajustam as análises anteriores. Por exemplo, para "GRUPO-CID-IIIA = SIM" os fatores A (tamanho da população), C (taxa de cruzamento) e D (taxa de mutação), assim como a interação entre os fatore AC e ACD, são os que mais influenciam significamente a variável de resposta e, portanto, pelo gráfico de efeitos principais definiu-se a intensidade dos ajustes desses parâmetros.

Tabela 6.8: Valores dos parâmetros para cada classe

Classe	**População**	**Gerações**	**Tx. Cruz.**	**Tx.Mut.**
GRUPO-CID-IIIA = Sim	100	70	75%	30%
GRUPO-CID-IIIB = Sim	100	60	90%	30%
GRUPO-EVENTO-09 = Sim	100	50	90%	30%

6.3.8 Resultados (PG)

Após a calibragem dos parâmetros utilizou-se o método de validação cruzada de k partições para avaliar o desempenho dos classificadores, no qual k (ver métodos de validação na Subseção 2.2.3) foi definido com valor 6 (empiricamente). A Tabela 6.9 apresenta o cálculo da média ($\bar{a}$) e desvio padrão (**DP**) da taxa de acerto e cobertura (*recall*) sobre o conjunto de teste das execuções em todas as partições para cada classe.

Tabela 6.9: Taxa de acerto utilizando validação cruzada para k = 6 (PG)

Atributo Meta	**Classe**	**Taxa Acerto** ($\bar{a}$%±**DP%**)	**Cobertura** ($\bar{a}$%±**DP%**)
GRUPO-CID-IIIA	Sim	34,5%±2,8%	1,80%±0,80%
GRUPO-CID-IIIB	Sim	5,9%±1,2%	0,76%±0,33%
GRUPO-EVENTO-09	Sim	81,3%±4,4%	0,58%±0,25%

6.3.9 Comparação de Resultados

Nesta subseção são apresentados os resultados da PG e do Algoritmo *C5.0* (Quinlan, 2011), com o objetivo de facilitar o entendimento dos resultados da PG. A PG foi executada também sem a restrição de tamanho[1] das regras uma vez que essa limitação não existia no *C5.0*, portanto por uma questão de simplicidade nomeou-se de PG1 a PG com restrições no tamanho das regras e PG2 a PG sem restrições.

Tabela 6.10: Taxa de acerto PG1 x PG2 x C5.0

Atributo Meta	**Classe**	**Tx. Ac. (PG1)** ($\bar{a}$%±**DP%**)	**Tx. Ac. (PG2)** ($\bar{a}$%±**DP%**)	**Tx. Ac. (C5.0)** ($\bar{a}$%±**DP%**)
GRUPO-CID-IIIA	Sim	34,5%±2,8%	33,0%±4,1%	**59,7%±8,2%**
GRUPO-CID-IIIB	Sim	5,9%±1,2%	10,9%±1,8%	**17,1%±13,3%**
GRUPO-EVENTO-09	Sim	**81,3%±4,4%**	74,7%±7,6%	75,4%±26,8%

Verifica-se pela Tabela 6.10 que a taxa de acerto é menor em PG para os atributos meta GRUPO-CID-IIIA e GRUPO-CID-IIIB e maior para GRUPO-EVENTO-09 para PG1 e similar para PG2, em relação ao *C5.0*.

Tabela 6.11: Cobertura PG1 x PG2 x C5.0

Atributo Meta	**Classe**	**Cober. (PG1)** ($\bar{a}$%±**DP%**)	**Cober. (PG2)** ($\bar{a}$%±**DP%**)	**Cober. (C5.0)** ($\bar{a}$%±**DP%**)
GRUPO-CID-IIIA	Sim	**1,80%±0,80%**	1,45%±0,64%	0,73%±0,14%
GRUPO-CID-IIIB	Sim	0,76%±0,33%	**0,84%±0,37%**	0,31%±0,09%
GRUPO-EVENTO-09	Sim	0,58%±0,25%	**1,00%±0,44%**	0,56±0,14

A Tabela 6.11 apresenta a média e o desvio padrão da cobertura das melhores regras encontradas, e pode-se verificar que a cobertura é maior nas duas abordagens de PG.

A Tabela 6.12 apresenta o tamanho médio dos antecedentes da melhor regra, e pode-se verificar que em PG1 o número de antecedentes é menor devido a restrição de tamanho das regras. O *C5.0* apesar de não ter restrição de tamanho das regras, em média não encontra regras muito grandes.

[1]Em PG foi definido que as regras deviam ter no máximo 5 antecedentes

Tabela 6.12: Média do número de antecedentes PG1 x PG2 x C5.0

Atributo Meta	**Classe**	**Média Antec. (PG1)** ($\bar{a}$)	**Média Antec. (PG2)** ($\bar{a}$)	**Média Antec. (C5.0)** ($\bar{a}$)
GRUPO-CID-IIIA	Sim	4,6	6,3	5
GRUPO-CID-IIIB	Sim	3,8	5,8	5
GRUPO-EVENTO-09	Sim	5	8	5,8

6.4 Segundo Experimento

O segundo experimento se refere a aplicação do algoritmo cultural apresentado no Capítulo 4 sobre os dados da saúde. Para realizar o experimento foi necessário obter as impressões gerais que foram fornecidas por um especialista em sistemas para planos de saúde. Para se obter essas impressões gerais foram apresentadas os atributos pré-selecionados no qual o especialista escolheu algumas e então criou as impressões gerais de acordo com o seu conhecimento sobre o tema. As impressões gerais são exibidas pela Tabela 6.13, 6.14 e 6.15.

Tabela 6.13: Impressões Gerais (GRUPO-CID-IIIA)

Num.	**Impressão Geral (IG)**
1	**SE** CONSULTA = ALTO **E** SADT = ALTO **ENTÃO** GRUPO-CID-IIIA = SIM
2	**SE** CONSULTA = ALTO **E** SADT = MEDIO **ENTÃO** GRUPO-CID-IIIA = SIM
3	**SE** ESPECIALIDADE-27 = SIM **E** SEXO = F **ENTÃO** GRUPO-CID-IIIA = SIM
4	**SE** ESPECIALIDADE-27 = SIM **E** SEXO = M **ENTÃO** GRUPO-CID-IIIA = SIM
5	**SE** IDADE = 60-A-69 **E** SEXO = F **ENTÃO** GRUPO-CID-IIIA = SIM
6	**SE** IDADE = 60-A-69 **E** SEXO = M **ENTÃO** GRUPO-CID-IIIA = SIM
7	**SE** IDADE = 70-A-79 **E** SEXO = F **ENTÃO** GRUPO-CID-IIIA = SIM
8	**SE** IDADE = 70-A-79 **E** SEXO = M **ENTÃO** GRUPO-CID-IIIA = SIM
9	**SE** IDADE = 80-OU-MAIS **E** SEXO = F **ENTÃO** GRUPO-CID-IIIA = SIM
10	**SE** IDADE = 80-OU-MAIS **E** SEXO = M **ENTÃO** GRUPO-CID-IIIA = SIM
11	**SE** GRUPO-EVENTO-01 = SIM **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIA = SIM
12	**SE** GRUPO-EVENTO-01 = SIM **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIA = SIM
13	**SE** GRUPO-EVENTO-04 = SIM **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIA = SIM
14	**SE** GRUPO-EVENTO-04 = SIM **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIA = SIM

Tabela 6.13: Impressões Gerais (GRUPO-CID-IIIA)

Num.	**Impressão Geral (IG)**
15	**SE** INTERNACAO = MAIS-QUE-UMA **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIA = SIM
16	**SE** INTERNACAO = MAIS-QUE-UMA **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIA = SIM

Tabela 6.14: Impressões Gerais (GRUPO-CID-IIIB)

Num.	**Impressão Geral (IG)**
1	**SE** CONSULTA = ALTO **E** SADT = ALTO **ENTÃO** GRUPO-CID-IIIB = SIM
2	**SE** CONSULTA = ALTO **E** SADT = MEDIO **ENTÃO** GRUPO-CID-IIIB = SIM
3	**SE** ESPECIALIDADE-27 = SIM **E** SEXO = F **ENTÃO** GRUPO-CID-IIIB = SIM
4	**SE** ESPECIALIDADE-27 = SIM **E** SEXO = M **ENTÃO** GRUPO-CID-IIIB = SIM
5	**SE** GRUPO-EVENTO-01 = SIM **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIB = SIM
6	**SE** GRUPO-EVENTO-01 = SIM **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIB = SIM
7	**SE** GRUPO-EVENTO-04 = SIM **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIB = SIM
8	**SE** GRUPO-EVENTO-04 = SIM **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIB = SIM
9	**SE** INTERNACAO = MAIS-QUE-UMA **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIB = SIM
10	**SE** INTERNACAO = MAIS-QUE-UMA **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIB = SIM
11	**SE** IDADE = 60-A-69 **E** SEXO = F **ENTÃO** GRUPO-CID-IIIB = SIM
12	**SE** IDADE = 60-A-69 **E** SEXO = M **ENTÃO** GRUPO-CID-IIIB = SIM
13	**SE** IDADE = 70-A-79 **E** SEXO = F **ENTÃO** GRUPO-CID-IIIB = SIM
14	**SE** IDADE = 70-A-79 **E** SEXO = M **ENTÃO** GRUPO-CID-IIIB = SIM
15	**SE** IDADE = 80-OU-MAIS **E** SEXO = F **ENTÃO** GRUPO-CID-IIIB = SIM
16	**SE** IDADE = 80-OU-MAIS **E** SEXO = M **ENTÃO** GRUPO-CID-IIIB = SIM

Tabela 6.15: Impressões Gerais (GRUPO-EVENTO-09)

Num.	**Impressão Geral (IG)**
1	**SE** INTERNACAO = UMA **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = SIM
2	**SE** INTERNACAO = MAIS-QUE-UMA **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = SIM
3	**SE** ESPECIALIDADE-64 = SIM **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = SIM

6.4.1 Experimento com interesse durante o processo evolutivo (AC1)

Neste experimento o cálculo do interesse das regras é feito durante o processo evolutivo, como citado na Seção 4.6.

Os experimentos foram realizados para os três atributos meta apresentados na Tabela 6.1, assim como no experimento em PG, mas considerando-se a classe "NAO"dos atributos, uma vez que as IG's fornecidas são da classe "SIM"e o interesse levado em consideração nesta dissertação são em regras com consequente contraditório, como discutido na subseção 4.6.2.

Para todos os casos foram utilizados os seguintes conjuntos de parâmetros determinados de maneira empírica:

1. Tamanho da população igual a 100;
2. Número de gerações igual a 100;
3. Número de indivíduos para elitismo igual a 2;
4. Taxa de cruzamento inicial igual a 50%;
5. Taxa de mutação inicial igual a 50%;
6. Taxa de aceitação igual a 10%;
7. Probabilidade inicial de influência do Conhecimento Situacional igual a 33,3%;
8. Probabilidade inicial de influência do Conhecimento Topográfico igual a 33,3%;
9. Probabilidade inicial de influência do Conhecimento de Domínio igual a 33,3%;
10. Número de melhores indivíduos armazenados no Conhecimento Situacional igual a 10;

Como descrito no capítulo anterior, os parâmetros de 4 a 9 são ajustados automaticamente ao longo da evolução, ao contrário da PG em que foi necessário criar um projeto de experimentos para ajustar os parâmetros.

Resultados

Utilizou-se o método de validação cruzada de k partes para avaliar o desempenho dos classificadores, no qual k foi definido com valor 6. A Tabela 6.16 apresenta o cálculo da média ($\bar{a}$) e desvio padrão (**DP**) da taxa de acerto, interesse e cobertura sobre o conjunto de teste da melhor regra do melhor indivíduo em cada execução da validação cruzada.

Tabela 6.16: Taxa de acerto utilizando validação cruzada para k = 6 (AC1)

Atributo Meta	**Classe**	**Taxa Acerto** ($\bar{a}$%±**DP%**)	**Interesse** ($\bar{a}$%±**DP%**)	**Cobertura** ($\bar{a}$%±**DP%**)
GRUPO-CID-IIIA	Não	99,0%±0,9%	100%±0,0%	0,10%±0,01%
GRUPO-CID-IIIB	Não	98,7%±0,4%	100%±0,0%	1,10%±0,03%
GRUPO-EVENTO-09	Não	93,2%±0,2%	100%±0,0%	4,69%±0,07%

Cabe salientar que nos experimentos realizados com o AC a variabilidade da melhor regra encontrada é menor, ou seja, em cada execução do algoritmo a melhor regra encontrada é na maioria das vezes a mesma, mesmo utilizando validação cruzada em que os conjuntos de treinamento em cada execução são ligeiramente diferentes, ao passo que em PG dificilmente a melhor regra de uma execução se repete. Entre os vários fatores que podem ter contribuído para isto pode-se destacar a utilização das impressões gerais pelo AC que guiam as soluções até uma mesma solução.

Outra questão importante é o tamanho das regras, pois a maioria das regras encontradas possuem apenas dois antecedentes. Um fator que contribuiu com o tamanho das regras encontradas foi o tamanho das IG's que possuem nesse experimento apenas dois antecedentes em que foram transformados em indivíduos (ver Seção 4.5.3) para influenciarem a evolução por meio do cruzamento. Para comprovar que o tamanho das IG's influenciam o tamanho das regras encontradas foram realizados experimentos simulando IG's com 5 antecedentes no qual obteve-se regras em média com 3 e 4 antecedentes.

No que se refere a evolução da taxa de acerto e interesse nas regras, os gráficos das Figuras 6.7 e 6.8 apresentam a evolução de uma das execuções para o atributo meta GRUPO-CID-IIIB. Percebe-se que desde a décima geração o algoritmo encontra regras com alta taxa de acerto. Já o interesse nas regras levam mais gerações para encontrar regras com interesse maior que 60%.

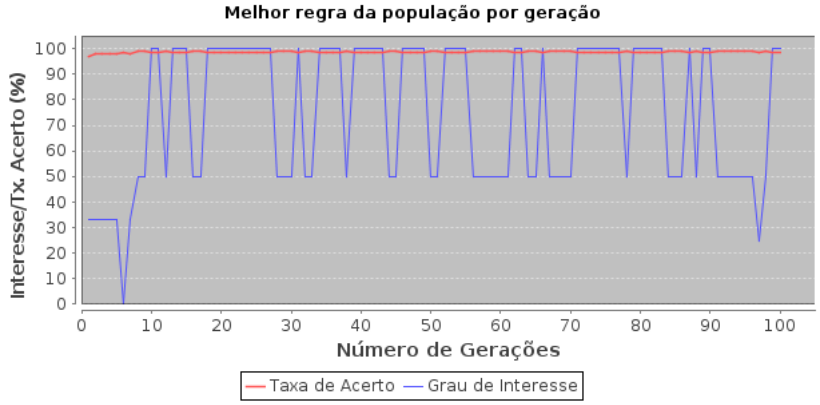

Figura 6.7: Evolução da taxa de acerto e interesse para GRUPO-CID-IIIB (AC1)

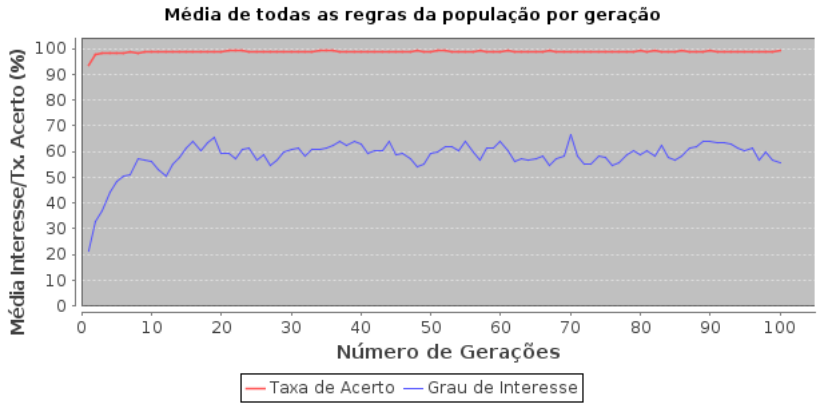

Figura 6.8: Evolução da média da taxa de acerto e interesse para GRUPO-CID--IIIB (AC1)

Os ajustes das taxas de mutação e cruzamento é sumarizada pela Figura 6.9. É possível observar que tanto a taxa de cruzamento quanto a taxa de mutação se estabilizam dentro de uma faixa de valores logo nas primeiras gerações.

A probabilidade de aplicação dos diferentes tipos de conhecimentos são mostrados na Figura 6.10. As funções de influência dos conhecimentos Situacional e Topográfico ocorrem a taxas maiores que a função de influência do Conhecimento de Domínio apesar de ao longo da evolução convergirem para valores próximos.

Os gráficos das Figuras 6.11 e 6.12 apresentam a evolução da aptidão em que é possível ver a rápida convergência das soluções.

Neste experimento, o AC1 conseguiu encontrar regras tanto pertencentes aos grandes disjuntos como aos pequenos disjuntos e com alta taxa de acerto, assim como regras com bom nível de interesse, no qual em alguns casos encontrou-se regras que contrariam totalmente as IG's do especialista de domínio. Por exemplo, foi encontrada a regra (**SE**

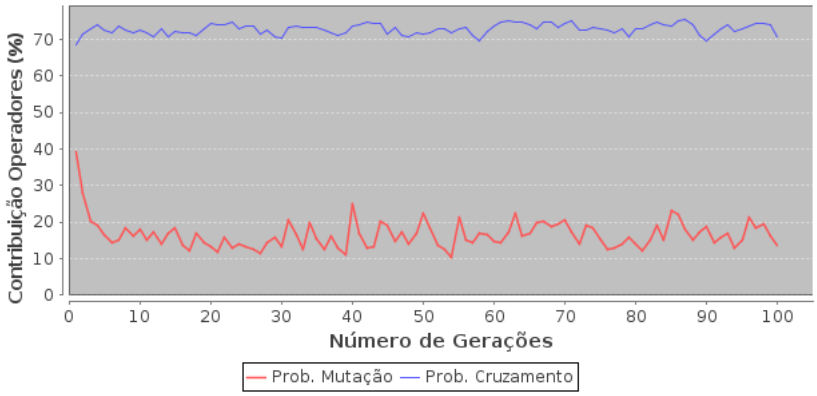

Figura 6.9: Ajuste das taxas de mutação e cruzamento para GRUPO-CID-IIIB (AC1) no decorrer do processo evolutivo

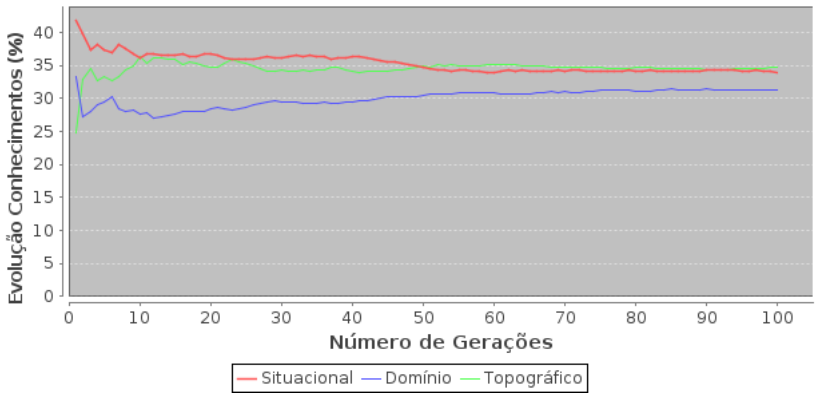

Figura 6.10: Ajuste das probabilidades dos conhecimentos para GRUPO-CID-IIIB (AC1) no decorrer do processo evolutivo

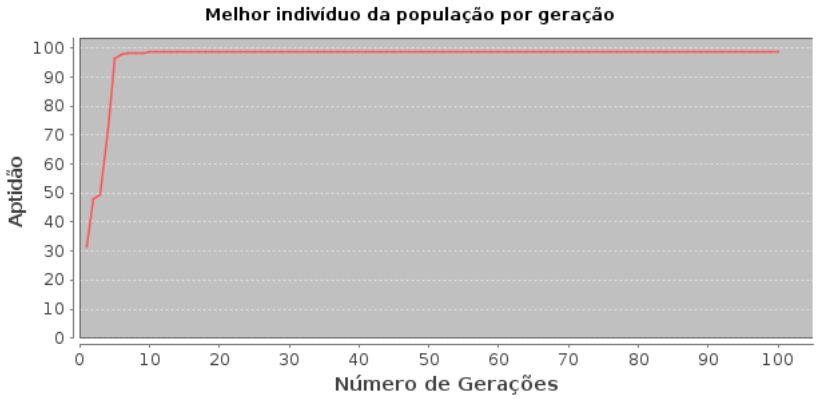

Figura 6.11: Evolução da aptidão para GRUPO-CID-IIIB (AC1)

GRUPO-CID-II = NAO **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIA = NAO)

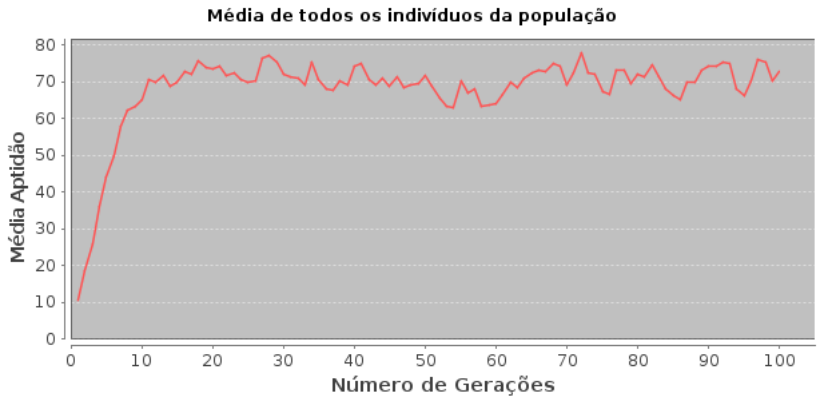

Figura 6.12: Evolução da média de aptidão da população para GRUPO-CID-IIIB (AC1)

na população final que possui 99,2% de taxa de acerto 45,57% de cobertura e interesse de 50%.

6.4.2 Experimento com interesse após o processo evolutivo (AC2)

Neste experimento o cálculo do interesse nas regras é feito após o processo evolutivo, como citado na seção 4.6. A única influência das IG's no processo evolutivo é por meio do conhecimento de domínio. Portanto o cálculo da aptidão do indivíduo é feito da mesma maneira que em PG (ver Equação 3.20).

Os experimentos foram realizados com os mesmos critérios do experimento da subseção 6.4.1 para os três atributos apresentados pela Tabela 6.4.1 e a classe "NAO".

Resultados

Utilizou-se o método de validação cruzada de k partes para avaliar o desempenho dos classificadores, no qual k foi definido com valor 6. A Tabela 6.17 apresenta o cálculo da média ($\bar{a}$) e desvio padrão (**DP**) da taxa de acerto, interesse e cobertura sobre o conjunto de teste da melhor regra de toda a população, ou seja, a regra que tem a maior taxa de acerto e interesse ao final de cada execução da validação cruzada.

Tabela 6.17: Taxa de acerto utilizando validação cruzada para k = 6 (AC2)

Atributo Meta	**Classe**	**Taxa Acerto** ($\bar{a}$%±**DP%**)	**Interesse** ($\bar{a}$%±**DP%**)	**Cobertura** ($\bar{a}$%±**DP%**)
GRUPO-CID-IIIA	Não	98,7%±0,6%	100,0%±0,0%	0,41%±0,19%
GRUPO-CID-IIIB	Não	98,9%±0,4%	94,4%±5,5%	0,64%±0,20%
GRUPO-EVENTO-09	Não	92,5%±0,3%	100,0%±0,0%	5,48%±0,78%

Nos experimentos realizados a variabilidade da melhor regra encontrada no AC2 é muito similar a abordagem da subseção 6.4.1 (AC1). Entretanto o tamanho médio das 10 melhores regras encontradas em AC2 é maior que em AC1.

No que se refere a evolução da taxa de acerto e interesse nas regras, os gráficos das Figuras 6.13 e 6.14 apresentam a evolução de uma das execuções para o atributo meta GRUPO-CID-IIIB. É possível notar que desde o início da execução o algoritmo encontra regras com alta taxa de acerto.

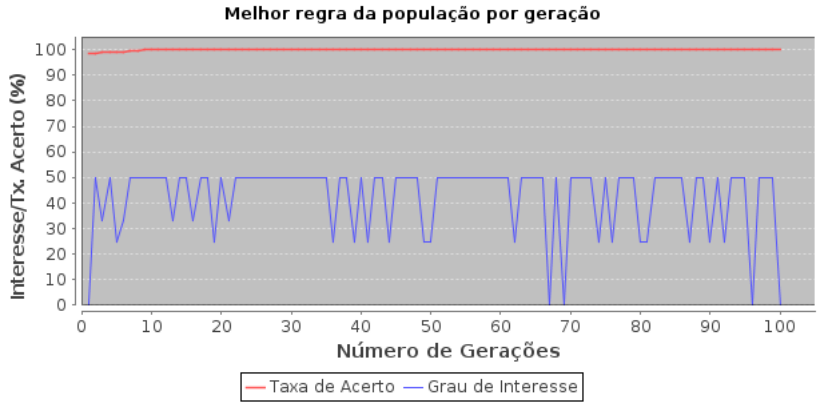

Figura 6.13: Evolução da taxa de acerto para GRUPO-CID-IIIB (AC2)

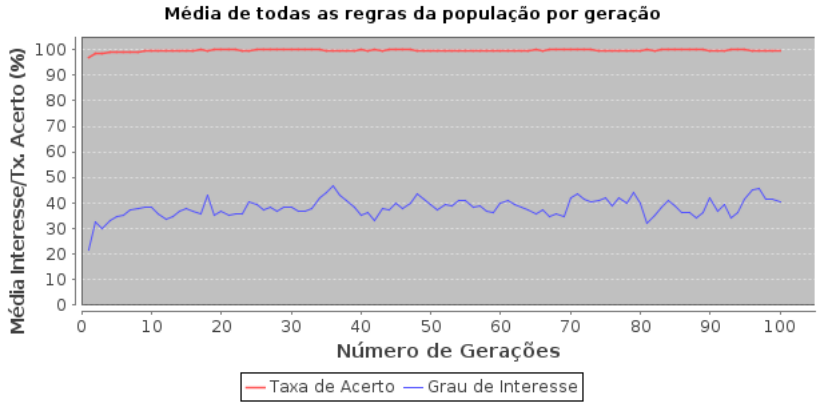

Figura 6.14: Evolução da média da taxa de acerto para GRUPO-CID-IIIB (AC2)

Os ajustes das taxas de mutação e cruzamento é sumarizada por meio do gráfico da Figura 6.15. É possível observar que a taxa de cruzamento se mantém relativamente estável enquanto a taxa de mutação sofre muita oscilação no decorrer das gerações.

A probabilidade de aplicação dos diferentes tipos de conhecimentos é mostrado na Figura 6.16. As funções de influência dos conhecimentos Situacional e Topográfico são aplicados a taxas menores que o conhecimento de domínio, o que mostra que a utilização

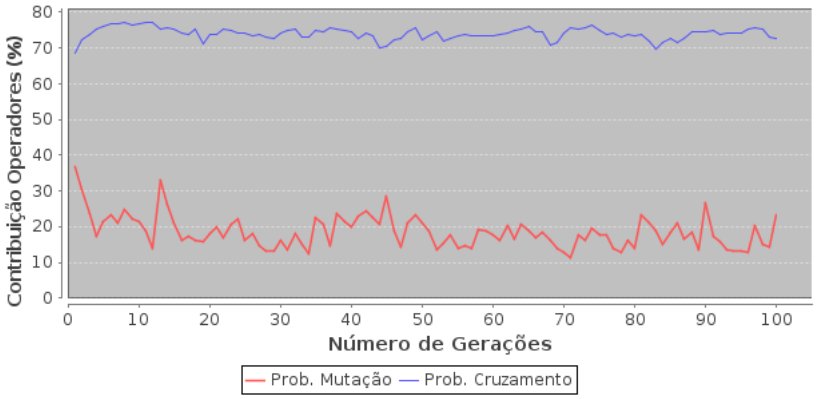

Figura 6.15: Ajuste das taxas de mutação e cruzamento para GRUPO-CID-IIIB (AC2) no decorrer do processo evolutivo

das IG's como conhecimento de domínio pode ser uma boa alternativa ao cálculo do interesse na função de aptidão, uma vez que aqui o número de cálculos pelo algoritmo é bastante reduzido.

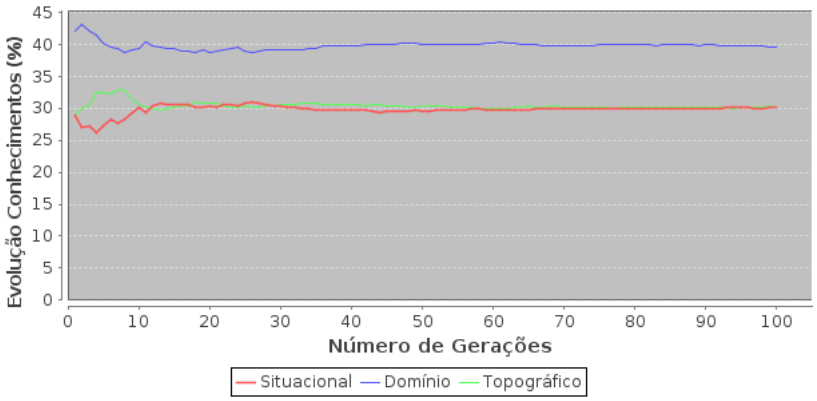

Figura 6.16: Ajuste das probabilidades dos conhecimentos para GRUPO-CID-IIIB (AC2) no decorrer do processo evolutivo

Os gráficos das Figuras 6.17 e 6.18 apresentam a evolução da aptidão em que é possível ver a rápida convergência das soluções.

6.4.3 Comparação de Resultados

Nesta subseção são apresentadas tabelas de maneira a simplificar a visualização dos resultados das duas abordagens do AC, e facilitar a comparação entre AC1 e o AC2.

Pela Tabela 6.18 é possível observar que os resultados são similares. A Tabela 6.19 mostra que o interesse nas regras encontradas também é similar para as duas abordagens de AC, indicando a escolha da abordagem de menor custo computacional.

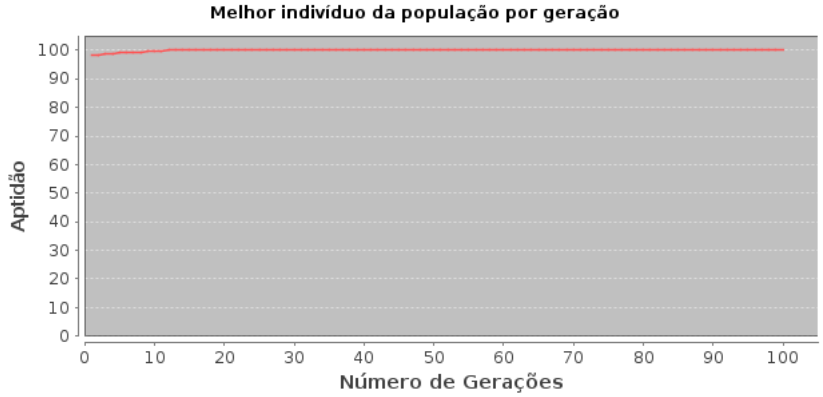

Figura 6.17: Evolução da aptidão para GRUPO-CID-IIIB (AC2)

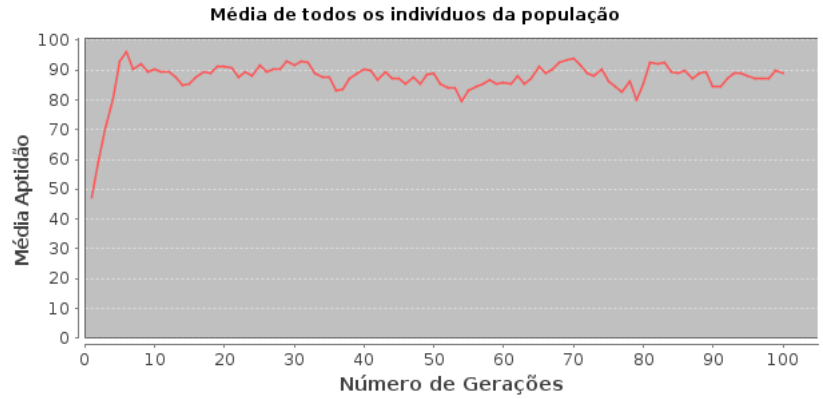

Figura 6.18: Evolução da média de aptidão para GRUPO-CID-IIIB (AC2)

Tabela 6.18: Taxa de acerto AC1 x AC2

Atributo Meta	**Classe**	**Taxa Acerto (AC1)** ($\bar{a}$%±**DP%**)	**Taxa Acerto (AC2)** ($\bar{a}$%±**DP%**)
GRUPO-CID-IIIA	Não	**99,0%±0,9%**	98,7%±0,6%
GRUPO-CID-IIIB	Não	98,7%±0,4%	**98,9%±0,4%**
GRUPO-EVENTO-09	Não	**93,2%±0,2%**	92,5%±0,3%

Tabela 6.19: Interesse AC1 x AC2

Atributo Meta	**Classe**	**Interesse (AC1)** ($\bar{a}$%±**DP%**)	**Interesse (AC2)** ($\bar{a}$%±**DP%**)
GRUPO-CID-IIIA	Não	100%±0,0%	100,0%±0,0%
GRUPO-CID-IIIB	Não	100%±0,0%	94,4%±5,5%
GRUPO-EVENTO-09	Não	100%±0,0%	100,0%±0,0%

Na Tabela 6.20 observa-se a cobertura média e o desvio padrão das duas abordagens, em que pode-se verificar que AC2 possui um desvio padrão maior.

Tabela 6.20: Cobertura AC1 x AC2

Atributo Meta	**Classe**	**Cobertura (AC1)** ($\bar{a}$%±**DP%**)	**Coberutra (AC2)** ($\bar{a}$%±**DP%**)
GRUPO-CID-IIIA	Não	0,10%±0,01%	**0,41%±0,19%**
GRUPO-CID-IIIB	Não	**1,10%±0,03%**	0,64%±0,20%
GRUPO-EVENTO-09	Não	4,69%±0,07%	**5,47%±0,78%**

Tabela 6.21: Média do número de antecedentes AC1 x AC2

Atributo Meta	**Classe**	**Média Antec. (AC1)** ($\bar{a}$)	**Média Antec. (AC2)** ($\bar{a}$)
GRUPO-CID-IIIA	Não	2,2	2,8
GRUPO-CID-IIIB	Não	2,0	2,7
GRUPO-EVENTO-09	Não	2,3	2,6

Em relação ao tamanho das regras encontradas é possível observar por meio da Tabela 6.21 que AC2 consegue regras maiores em relação a AC1.

6.5 Regras Encontradas

Nesta seção são apresentadas as 10 melhores regras diferentes para cada experimento executado anteriormente para cada atributo meta e classe.

6.5.1 Regras PG

Tabela 6.22: Regras para o atributo GRUPO-CID-IIIA (PG)

Regra	**Taxa Acerto (%)**	**Cobert. (*Recall*) (%)**
SE ESPECIALIDADE-04 = SIM **E** CONSULTA = ALTO **E** ESPECIALIDADE-39 = NAO **E** ESPECIALIDADE-27 = SIM **E** GRUPO-CID-IIIB = NAO **ENTÃO** GRUPO-CID-IIIA = SIM	40,4%	1,88%
SE ESPECIALIDADE-04 = SIM **E** CONSULTA = ALTO **E** GRUPO-CID-VIIA = NAO **E** ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-39 = NAO **ENTÃO** GRUPO-CID-IIIA = SIM	36,5%	1,48%
SE ESPECIALIDADE-04 = SIM **E** ESPECIALIDADE-39 = NAO **E** CONSULTA = ALTO **E** ESPECIALIDADE-27 = SIM **E** GRUPO-CID-XB = NAO **ENTÃO** GRUPO-CID-IIIA = SIM	41,1%	2,07%

Tabela 6.22: Regras para o atributo GRUPO-CID-IIIA (PG)

Regra	**Taxa Acerto (%)**	**Cobert. (*Recall*) (%)**
SE ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-04 = SIM **E** GRUPO-CID-VIIA = NAO **E** CONSULTA = ALTO **E** ESPECIALIDADE-13 = NAO **ENTÃO** GRUPO-CID-IIIA = SIM	38,7%	1,88%
SE ESPECIALIDADE-04 = SIM **E** CONSULTA = ALTO **E** ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-39 = NAO **ENTÃO** GRUPO-CID-IIIA = SIM	39,6%	2,07%
SE ESPECIALIDADE-04 = SIM **E** CONSULTA = ALTO **E** ESPECIALIDADE-27 = SIM **E** GRUPO-CID-IIIB = NAO **E** ESPECIALIDADE-54 = NAO **ENTÃO** GRUPO-CID-IIIA = SIM	38,1%	2,07%
SE CONSULTA = ALTO **E** GRUPO-CID-IIIB = NAO **E** ESPECIALIDADE-04 = SIM **E** ESPECIALIDADE-27 = SIM **ENTÃO** GRUPO-CID-IIIA = SIM	37,7%	2,27%
SE ESPECIALIDADE-04 = SIM **E** CONSULTA = ALTO **E** ESPECIALIDADE-39 = NAO **E** ESPECIALIDADE-27 = SIM **E** GRUPO-CID-XII = NAO **ENTÃO** GRUPO-CID-IIIA = SIM	37,2%	1,88%
SE ESPECIALIDADE-04 = SIM **E** GRUPO-CID-VIIA = NAO **E** GRUPO-CID-XB = NAO **E** CONSULTA = ALTO **E** ESPECIALIDADE-27 = SIM **ENTÃO** GRUPO-CID-IIIA = SIM	35,8%	1,88%
SE ESPECIALIDADE-04 = SIM **E** ESPECIALIDADE-27 = SIM **E** CONSULTA = ALTO **E** ESPECIALIDADE-58 = NAO **ENTÃO** GRUPO-CID-IIIA = SIM	41,5%	2,67%

Tabela 6.23: Regras para o atributo GRUPO-CID-IIIB (PG)

Regra	**Taxa Acerto (%)**	**Cobert. (*Recall*) (%)**
SE GRUPO-CID-VIB = NAO **E** ESPECIALIDADE-27 = SIM **E** GRUPO-CID-XIII = NAO **E** ESPECIALIDADE-53 = NAO **E** GRUPO-CID-XVI = SIM **ENTÃO** GRUPO-CID-IIIB = SIM	8,6%	0,93%
SE GRUPO-CID-XVI = SIM **E** ESPECIALIDADE-27 = SIM **E** GRUPO-CID-VIB = NAO **E** ESPECIALIDADE-64 = NAO **E** ESPECIALIDADE-53 = NAO **ENTÃO** GRUPO-CID-IIIB = SIM	11,9%	1,17%

Tabela 6.23: Regras para o atributo GRUPO-CID-IIIB (PG)

Regra	**Taxa Acerto (%)**	**Cobert. (*Recall*) (%)**
SE GRUPO-CID-XIII = NAO **E** ESPECIALIDADE-53 = NAO **E** GRUPO-CID-XVI = SIM **E** ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-64 = NAO **ENTÃO** GRUPO-CID-IIIB = SIM	9,5%	0,93%
SE GRUPO-CID-VIB = NAO **E** ESPECIALIDADE-53 = NAO **E** ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-24 = NAO **E** GRUPO-CID-XVI = SIM **ENTÃO** GRUPO-CID-IIIB = SIM	9,8%	1,17%
SE ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-53 = NAO **E** GRUPO-CID-XVI = SIM **E** GRUPO-CID-XIII = NAO **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIB = SIM	9,2%	1,17%
SE ESPECIALIDADE-27 = SIM **E** GRUPO-CID-XIII = NAO **E** ESPECIALIDADE-53 = NAO **E** GRUPO-CID-XVI = SIM **ENTÃO** GRUPO-CID-IIIB = SIM	9,2%	1,17%
SE GRUPO-CID-XVI = SIM **E** ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-64 = NAO **E** CONSULTA = ALTO **E** ESPECIALIDADE-53 = NAO **ENTÃO** GRUPO-CID-IIIB = SIM	12,5%	1,40%
SE CONSULTA = ALTO **E** ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-24 = NAO **E** ESPECIALIDADE-53 = NAO **E** GRUPO-CID-XVI = SIM **ENTÃO** GRUPO-CID-IIIB = SIM	10,0%	1,40%
SE CONSULTA = ALTO **E** GRUPO-CID-XVI = SIM **E** ESPECIALIDADE-47 = NAO **E** ESPECIALIDADE-27 = SIM **E** GRUPO-CID-VIB = NAO **ENTÃO** GRUPO-CID-IIIB = SIM	8,0%	1,17%
SE ESPECIALIDADE-27 = SIM **E** ESPECIALIDADE-24 = NAO **E** GRUPO-CID-XVI = SIM **E** ESPECIALIDADE-47 = NAO **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIB = SIM	8,0%	1,17%

Tabela 6.24: Regras para o atributo GRUPO-EVENTO-09 (PG)

Regra	**Taxa Acerto (%)**	**Cobert. (*Recall*) (%)**
SE ESPECIALIDADE-64 = SIM **E** IDADE = 40-A-49 **E** ESPECIALIDADE-33 = SIM **E** ESPECIALIDADE-56 = SIM **E** CONSULTA = MEDIO **ENTÃO** GRUPO-EVENTO-09 = SIM	100,0%	0,38%

Tabela 6.24: Regras para o atributo GRUPO-EVENTO-09 (PG)

Regra	**Taxa Acerto (%)**	**Cobert. (*Recall*) (%)**
SE ESPECIALIDADE-64 = SIM **E** IDADE = 40-A-49 **E** ESPECIALIDADE-33 = SIM **E** GRUPO-EVENTO-04 = NAO **E** CONSULTA = MEDIO **ENTÃO** GRUPO-EVENTO-09 = SIM	89,4%	1,10%
SE ESPECIALIDADE-56 = SIM **E** ESPECIALIDADE-47 = SIM **E** ESPECIALIDADE-33 = SIM **E** ESPECIALIDADE-64 = SIM **ENTÃO** GRUPO-EVENTO-09 = SIM	63,6%	0,90%
SE ESPECIALIDADE-56 = SIM **E** ESPECIALIDADE-33 = SIM **E** ESPECIALIDADE-64 = SIM **E** ESPECIALIDADE-09 = NAO **E** IDADE = 40-A-49 **ENTÃO** GRUPO-EVENTO-09 = SIM	75,0%	2,72%
SE ESPECIALIDADE-69 = NAO **E** IDADE = 40-A-49 **E** ESPECIALIDADE-33 = SIM **E** ESPECIALIDADE-64 = SIM **E** ESPECIALIDADE-56 = SIM **ENTÃO** GRUPO-EVENTO-09 = SIM	78,3%	3,75%
SE ESPECIALIDADE-33 = SIM **E** CONSULTA = MEDIO **E** ESPECIALIDADE-64 = SIM **E** ESPECIALIDADE-59 = NAO **E** ESPECIALIDADE-56 = SIM **ENTÃO** GRUPO-EVENTO-09 = SIM	67,5%	1,74%
SE ESPECIALIDADE-64 = SIM **E** IDADE = 40-A-49 **E** ESPECIALIDADE-33 = SIM **E** ESPECIALIDADE-56 = SIM **E** ESPECIALIDADE-59 = NAO **ENTÃO** GRUPO-EVENTO-09 = SIM	77,1%	3,49%
SE ESPECIALIDADE-33 = SIM **E** CONSULTA = MEDIO **E** ESPECIALIDADE-64 = SIM **E** ESPECIALIDADE-56 = SIM **ENTÃO** GRUPO-EVENTO-09 = SIM	67,5%	1,74%
SE ESPECIALIDADE-64 = SIM **E** IDADE = 40-A-49 **E** ESPECIALIDADE-27 = NAO **E** ESPECIALIDADE-56 = SIM **E** ESPECIALIDADE-33 = SIM **ENTÃO** GRUPO-EVENTO-09 = SIM	81,0%	3,88%
SE ESPECIALIDADE-56 = SIM **E** ESPECIALIDADE-64 = SIM **E** SADT = ALTO **E** ESPECIALIDADE-33 = SIM **ENTÃO** GRUPO-EVENTO-09 = SIM	60,0%	2,13%

6.5.2 Regras AC1

Tabela 6.25: Regras para o atributo GRUPO-CID-IIIA (AC1)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE GRUPO-EVENTO-01 = SIM **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIA = NAO	100,0%	100,0%	0,12%
SE GRUPO-EVENTO-04 = SIM **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIA = NAO	100,0%	100,0%	0,05%
SE INTERNACAO = MAIS-QUE-UMA **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIA = NAO	98,9%	100,0%	1,08%
SE SEXO = M **E** IDADE = 60-A-69 **ENTÃO** GRUPO-CID-IIIA = NAO	96,5%	100,0%	4,29%
SE SEXO = F **E** IDADE = 80-OU-MAIS **ENTÃO** GRUPO-CID-IIIA = NAO	95,6%	100,0%	1,49%
SE IDADE = 60-A-69 **E** SEXO = F **ENTÃO** GRUPO-CID-IIIA = NAO	96,1%	100,0%	6,06%
SE CONSULTA = ALTO **E** SADT = MEDIO **ENTÃO** GRUPO-CID-IIIA = NAO	95,1%	100,0%	14,13%
SE SEXO = F **E** ESPECIALIDADE-27 = SIM **ENTÃO** GRUPO-CID-IIIA = NAO	87,3%	100,0%	1,30%
SE GRUPO-EVENTO-04 = SIM **E** ESPECIALIDADE-66 = NAO **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIA = NAO	93,5%	66,7%	0,38%
SE GRUPO-EVENTO-01 = SIM **E** CONSULTA = ALTO **E** SEXO = M **ENTÃO** GRUPO-CID-IIIA = NAO	96,2%	66,7%	0,11%

Tabela 6.26: Regras para o atributo GRUPO-CID-IIIB (AC1)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE INTERNACAO = MAIS-QUE-UMA **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIB = NAO	99,4%	100,0%	1,13%

Tabela 6.26: Regras para o atributo GRUPO-CID-IIIB (AC1)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE SEXO = M **E** IDADE = 70-A-79 **ENTÃO** GRUPO-CID-IIIB = NAO	98,4%	100,0%	1,99%
SE GRUPO-EVENTO-04 = SIM **E** CONSULTA = ALTO **ENTÃO** GRUPO-CID-IIIB = NAO	97,6%	100,0%	0,37%
SE SEXO = F **E** IDADE = 80-OU-MAIS **ENTÃO** GRUPO-CID-IIIB = NAO	99,3%	100,0%	1,33%
SE SEXO = M **E** IDADE = 80-OU-MAIS **ENTÃO** GRUPO-CID-IIIB = NAO	99,2%	100,0%	0,61%
SE SEXO = M **E** IDADE = 60-A-69 **ENTÃO** GRUPO-CID-IIIB = NAO	97,7%	100,0%	4,11%
SE SEXO = F **E** IDADE = 70-A-79 **ENTÃO** GRUPO-CID-IIIB = NAO	97,2%	100,0%	3,13%
SE CONSULTA = ALTO **E** INTERNACAO = MAIS-QUE-UMA **ENTÃO** GRUPO-CID-IIIB = NAO	95,6%	100,0%	6,36%
SE SEXO = M **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIB = NAO	99,3%	50,0%	22,96%
SE CONSULTA = MEDIO **E** ESPECIALIDADE-31 = SIM **ENTÃO** GRUPO-CID-IIIB = NAO	100,0%	50,0%	0,33%

Tabela 6.27: Regras para o atributo GRUPO-EVENTO-09 (AC1)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE INTERNACAO = MAIS-QUE-UMA **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = NAO	93,7%	100,0%	4,93%
SE INTERNACAO = UMA **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = NAO	92,0%	100,0%	8,86%
SE ESPECIALIDADE-64 = SIM **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = NAO	68,3%	100,0%	2,49%
SE ESTADO = PR **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = NAO	94,0%	50,0%	0,28%

Tabela 6.27: Regras para o atributo GRUPO-EVENTO-09 (AC1)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE INTERNACAO = MAIS-QUE-UMA **E** GRUPO-CID-XII = SIM **ENTÃO** GRUPO-EVENTO-09 = NAO	93,8%	50,0%	0,34%
SE ESPECIALIDADE-05 = NAO **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = NAO	93,7%	50,0%	56,55%
SE INTERNACAO = MAIS-QUE-UMA **E** ESPECIALIDADE-53 = SIM **ENTÃO** GRUPO-EVENTO-09 = NAO	94,3%	50,0%	0,60%
SE GRUPO-EVENTO-01 = NAO **E** ESPECIALIDADE-33 = NAO **E** ESPECIALIDADE-64 = SIM **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = NAO	83,4%	50,0%	1,81%
SE ESPECIALIDADE-64 = SIM **E** ESPECIALIDADE-59 = NAO **ENTÃO** GRUPO-EVENTO-09 = NAO	76,9%	50,0%	3,80%
SE ESPECIALIDADE-57 = SIM **E** SEXO = F **E** INTERNACAO = MAIS-QUE-UMA **ENTÃO** GRUPO-EVENTO-09 = NAO	95,8%	66,7%	0,74%

6.5.3 Regras AC2

Tabela 6.28: Regras para o atributo GRUPO-CID-IIIA (AC2)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE CONSULTA = MEDIO **E** INTERNACAO = MAIS-QUE-UMA **ENTÃO** GRUPO-CID-IIIA = NAO	98,4%	100,0%	0,96%
SE CONSULTA = ALTO **E** SADT = MEDIO **ENTÃO** GRUPO-CID-IIIA = NAO	95,6%	100,0%	16,67%
SE SEXO = F **E** IDADE = 70-A-79 **ENTÃO** GRUPO-CID-IIIA = NAO	96,2%	100,0%	3,24%

Tabela 6.28: Regras para o atributo GRUPO-CID-IIIA (AC2)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE GRUPO-CID-VIB = SIM **E** CONSULTA = MEDIO **E** INTERNACAO = MAIS-QUE-UMA **ENTÃO** GRUPO-CID-IIIA = NAO	100,0%	66,7%	0,04%
SE CONSULTA = MEDIO **E** SEXO = F **E** IDADE = 60-A-69 **ENTÃO** GRUPO-CID-IIIA = NAO	99,0%	66,7%	1,83%
SE SEXO = M **E** INTERNACAO = MAIS-QUE-UMA **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIA = NAO	96,9%	66,7%	0,49%
SE SEXO = M **E** CONSULTA = ALTO **E** GRUPO-EVENTO-04 = SIM **ENTÃO** GRUPO-CID-IIIA = NAO	90,4%	66,7%	0,04%
SE IDADE = 60-A-69 **E** SEXO = M **E** GRUPO-CID-VIB = SIM **ENTÃO** GRUPO-CID-IIIA = NAO	98,1%	66,7%	0,24%
SE SEXO = M **E** IDADE = 80-OU-MAIS **E** GRUPO-CID-VIB = SIM **ENTÃO** GRUPO-CID-IIIA = NAO	91,6%	66,7%	0,04%
SE SEXO = M **E** CONSULTA = MEDIO **E** IDADE = 60-A-69 **E** GRUPO-CID-VIIIA = SIM **ENTÃO** GRUPO-CID-IIIA = NAO	100,0%	50,0%	0,01%

Tabela 6.29: Regras para o atributo GRUPO-CID-IIIB (AC2)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE CONSULTA = MEDIO **E** INTERNACAO = MAIS-QUE-UMA **ENTÃO** GRUPO-CID-IIIB = NAO	99,7%	100,0%	1,10%
SE IDADE = 80-OU-MAIS **E** SEXO = M **ENTÃO** GRUPO-CID-IIIB = NAO	98,0%	100,0%	0,68%
SE CONSULTA = ALTO **E** GRUPO-EVENTO-01 = SIM **ENTÃO** GRUPO-CID-IIIB = NAO	96,8%	100,0%	0,55%

Tabela 6.29: Regras para o atributo GRUPO-CID-IIIB (AC2)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE GRUPO-EVENTO-04 = SIM **E** CONSULTA = MEDIO **ENTÃO** GRUPO-CID-IIIB = NAO	91,6%	100,0%	0,04%
SE ESPECIALIDADE-27 = SIM **E** CONSULTA = MEDIO **E** SEXO = M **ENTÃO** GRUPO-CID-IIIB = NAO	100,0%	66,7%	0,13%
SE GRUPO-CID-XB = NAO **E** CONSULTA = ALTO **E** GRUPO-EVENTO-01 = SIM **ENTÃO** GRUPO-CID-IIIB = NAO	96,8%	66,7%	0,54%
SE GRUPO-EVENTO-01 = SIM **E** CONSULTA = ALTO **E** GRUPO-CID-II = SIM **ENTÃO** GRUPO-CID-IIIB = NAO	100,0%	66,7%	0,04%
SE IDADE = 60-A-69 **E** CONSULTA = MEDIO **E** SEXO = M **E** GRUPO-CID-II = SIM **ENTÃO** GRUPO-CID-IIIB = NAO	100,0%	50,0%	0,04%
SE ESPECIALIDADE-59 = NAO **E** CONSULTA = MEDIO **E** SEXO = M **E** GRUPO-EVENTO-04 = SIM **ENTÃO** GRUPO-CID-IIIB = NAO	100,0%	50,0%	0,02%
SE CONSULTA = MEDIO **E** SEXO = M **ENTÃO** GRUPO-CID-IIIB = NAO	99,0%	50,0%	23,07%

Tabela 6.30: Regras para o atributo GRUPO-EVENTO-09 (AC2)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE INTERNACAO = MAIS-QUE-UMA **E** SEXO = F **ENTÃO** GRUPO-EVENTO-09 = NAO	93,7%	100,0%	4,93%
SE SEXO = F **E** ESPECIALIDADE-64 = SIM **ENTÃO** GRUPO-EVENTO-09 = NAO	68,3%	100,0%	2,49%
SE SEXO = F **E** IDADE = 10-A-19 **E** ESPECIALIDADE-64 = SIM **ENTÃO** GRUPO-EVENTO-09 = NAO	100,0%	66,7%	0,16%

Tabela 6.30: Regras para o atributo GRUPO-EVENTO-09 (AC2)

Regra	**Tx. Ac.** (%)	**Inter.** (%)	**Cobert.** (***Recall***) (%)
SE SEXO = M **E** INTERNACAO = UMA **ENTÃO** GRUPO-EVENTO-09 = NAO	99,7%	50,0%	5,25%
SE INTERNACAO = MAIS-QUE-UMA **E** ESPECIALIDADE-47 = SIM **ENTÃO** GRUPO-EVENTO-09 = NAO	94,6%	50,0%	0,48%
SE ESPECIALIDADE-64 = SIM **E** GRUPO-CID-XB = SIM **ENTÃO** GRUPO-EVENTO-09 = NAO	83,7%	50,0%	0,14%
SE INTERNACAO = MAIS-QUE-UMA **E** MUNICIPIO = PALMITOS **ENTÃO** GRUPO-EVENTO-09 = NAO	100,0%	50,0%	0,04%
SE ESPECIALIDADE-64 = SIM **E** GRUPO-CID-XA = NAO **E** SEXO = F **E** IDADE = 10-A-19 **ENTÃO** GRUPO-EVENTO-09 = NAO	100,0%	50,0%	0,16%
SE SADT = BAIXO **E** IDADE = 10-A-19 **E** ESPECIALIDADE-64 = SIM **ENTÃO** GRUPO-EVENTO-09 = NAO	100,0%	33,3%	0,21%
SE GRUPO-CID-XA = NAO **E** ESPECIALIDADE-69 = NAO **E** INTERNACAO = UMA **E** ESPECIALIDADE-47 = SIM **ENTÃO** GRUPO-EVENTO-09 = NAO	89,3%	25,0%	0,38%

6.6 Considerações Finais

Os experimentos apresentados neste capítulo teve como objetivo avaliar o algoritmo de programação genética e o algoritmo cultural implementados. O algoritmo de programação genética obteve resultados satisfatórios quando comparado com o algoritmo C5.0 e mostrou-se interessante para o objetivo pelo qual foi implementado que é ser utilizado como componente populacional (núcleo) do algoritmo cultural.

Por meio desses experimentos foi possível observar a influência dos tipos de conhecimentos implementados no algoritmo cultural, assim como a utilização das impressões gerais como conhecimento de domínio. Foi possível observar também a influência da

utilização do conhecimento de domínio em detrimento do cálculo do grau de interesse na função de aptidão.

Outro fator importante observado por meio desses experimentos e resultados foi a utilização do ajuste automático de parâmetros, que mostrou-se útil para mineração de dados.

Capítulo

7

Conclusões e Trabalhos Futuros

7.1 Conclusões

Nesta dissertação apresentou-se uma abordagem de algoritmos culturais para a tarefa de classificação em mineração de dados, utilizando como componente populacional um algoritmo de programação genética. Os resultados obtidos comprovam a capacidade do algoritmo em encontrar regras com alta taxa de acerto, incluindo regras pertencentes aos grandes disjuntos como aos pequenos disjuntos. Além disso, a utilização de lógica difusa simplificou a etapa de preparação de dados evitando o tratamento de atributos contínuos.

Como estudo de caso, utilizou-se dados de um plano de saúde suplementar que passaram por uma etapa de preparação. Nessa preparação obteve-se uma redução de 120 para 55 atributos e redução em 14% no número de registros considerados. Obteve-se também dados de melhor qualidade e adequados para utilização por algoritmos de MD. Esses dados ainda podem sofrer algum processo de refinamento dependendo do foco que se queira dar em sua utilização futura.

Esta dissertação também apresentou uma abordagem para utilização de impressões gerais como um dos conhecimentos do algoritmo cultural, mais especificamente a utilização de impressões gerais como conhecimento de domínio do algoritmo cultural, que mostrou-se interessante, pois o algoritmo proposto inicialmente foi executado (AC1) considerando as impressões gerais durante o processo evolutivo por meio do cálculo do interesse nas regras geradas. Em seguida realizou-se outro experimento (AC2) em que o cálculo do grau de

interesse é feito após o processo evolutivo, e os resultados do AC1 e AC2 não mostraram diferenças significativas.

Considerando os três tipos de conhecimentos (situacional, topográfico e de domínio), todos apresentaram desempenho similar em suas utilizações sendo que o conhecimento de domínio se destacou em alguns experimentos o que reforça que a utilização do conhecimento de domínio é uma boa alternativa para influência das impressões gerais sobre a população.

O algoritmo cultural proposto possui alguns pontos de adaptação como as taxas de aplicação das funções de influência, probabilidade de uso da mutação ou cruzamento e taxa de aceitação. Acredita-se que as adaptações automáticas de parâmetros propostas mostram-se eficientes, visto que os resultados obtidos foram bons, além de ser uma técnica importante para o ajuste de parâmetros em MD, pois conforme se altera os dados utilizados para mineração dificilmente os parâmetros do algoritmo serão os mesmos para se obter boas soluções.

As regras encontradas pelo algoritmo cultural apresentam boa taxa de acerto e bom nível de interesse. Infelizmente não foi possível fazer comparações diretas com outros algoritmos devido a dificuldade em encontrar implementações que utilizem impressões gerais. Entretanto, a maioria das regras possuem boa taxa de acerto e várias coberturas o que indica que o algoritmo proposto encontra regras tanto dos pequenos disjuntos como dos grandes disjuntos.

7.2 Trabalhos Futuros

Como principais trabalhos futuros propõe-se:

- Desenvolver um algoritmo cultural para classificação em mineração de dados utilizando outros algoritmos como componente populacional (núcleo).
- Executar o algoritmo proposto nesta dissertação considerando outras impressões gerais e outros atributos metas, buscando encontrar mais regras para dados da saúde.
- Fazer comparações diretas do algoritmo proposto com outras abordagens que utilizem impressões gerais.
- Desenvolver outros tipos de conhecimentos para o algoritmo cultural para mineração de dados.

- Desenvolver um conhecimento de domínio para o algoritmo cultural que utilize métodos *data-driven* em vez de *user-driven*.
- Aplicar o AC sobre outras bases de dados reais.
- Avaliar a informação descoberta por meio de questionário junto ao usuário para medir o real interesse.

Referências

Agência nacional de saúde suplementar. Acesso em 15 dez. 2011, 2011. Disponível em `http://www.ans.gov.br/index.php/component/content/article/65/598-guia-de-servicos-profissionaisservico-auxiliar-diagnostico-e-terapia`

Ministério da saúde - departamento de informática do sus. Acesso em 15 dez. 2011, 2011. Disponível em `http://www.datasus.gov.br/cid10/v2008/webhelp/z00_z13.htm#Z00`

Abarghouei, A.; Ghanizadeh, A.; Shamsuddin, S. Advances of Soft Computing Methods in Edge Detection. *Int. J. Advance. Soft Comput. Appl*, v. 1, n. 2, 2009.

Back, T.; Fogel, D. B.; Michalewicz, Z., eds. *Basic algorithms and operators*. 1st ed. Bristol, UK, UK: IOP Publishing Ltd., 1999.

Banzhaf, W.; Peter, N.; Robert, E., K.; Frank, D., F. *Genetic Programming: An Introduction on the Automatic Evolution of computer programs and its Applications*. Morgan Kaufmann, 1998.

Batista, G. *Pré-processamento de dados em aprendizado de máquina supervisionado*. Tese de Doutoramento, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2003.

Becerra, R. *Algoritmos culturales aplicados a optimización con restricciones y optimización multiobjetivo*. Tese de Doutoramento, Instituto Politécnico Nacional do México, 2002.

Becerra, R.; Coello, C. Optimization with Constraints using a Cultured Differential Evolution Approach. *optimization*, v. 22, n. 4, p. 11–24, 2005.

Bloedorn, E.; Michalsi, R. Data-driven constructive induction. *Intelligent Systems and their Applications, IEEE*, v. 13, n. 2, p. 30 –37, 1998.

Bojarczuk, C.; Lopes, H.; Freitas, A. Discovering comprehensible classification rules using genetic programming: a case study in a medical domain. In: *Proceedings of the Genetic and Evolutionary Computation Conference*, Citeseer, 1999, p. 953–958.

Bojarczuk, C.; Lopes, H.; Freitas, A. Genetic programming for knowledge discovery in chest-pain diagnosis. *Engineering in Medicine and Biology Magazine, IEEE*, v. 19, n. 4, p. 38 –44, 2000.

Bojarczuk, C. C.; Lopes, H. S.; Freitas, A. A.; Michalkiewicz, E. L. A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets. *Artificial Intelligence in Medicine*, v. 30, n. 1, p. 27 – 48, 2004.

Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Classification and regression trees. 1984. *Wadsworth, Belmont, CA*, v. 1, 1984.

Carneiro, M.; Araújo, S.; Lemos, R. Programação genética aplicada à linguagem ladder. In: *Induscon–VIII Conferência Internacional de Aplicações Industriais, Poços de Caldas, Minas Gerais, Brasil*, 2008.

Carvalho, D. R. *Árvore de decisão / algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados*. Tese de Doutoramento, COPPE/UFRJ, 2005.

de Castro, M. S. M.; Carvalho, M. S. Agrupamento da classificação internancional de doenças para análise de reinternação hospitalares. *Cad. Saúde Pública*, v. 21(1), p. 317–323, 2005.

Cavalcante, P. F. *A importância dos fatores de risco na obstrução das artérias coronárias utilizando técnicas de mineração de dados*. Dissertação de Mestrado, Universidade Católica de Goiás, 2009.

Cavaretta, M. J.; Chellapilla, K. Data mining using genetic programming: The implications of parsimony on generalization error. In: *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99*, Citeseer, 1999, p. 1330–1337.

Chen, M.; Han, J.; Yu, P. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, v. 8, n. 6, p. 866–883, 1996.

Coelho, R. C. *Um algoritmo genético difuso multiobjetivo para descoberta de conhecimento*. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual De Maringá, 2004.

Colorni, A.; Dorigo, M.; Maniezzo, V.; *et al.* Distributed optimization by ant colonies. In: *Proceedings of the First European Conference on Artificial Life*, 1992, p. 134–142.

Dallagassa, M.; Silva, S.; Carvalho, D. Avaliação de modelos para a classificação de beneficiários com indicativos para o diabetes mellitus tipo 2. *Instituto Paranaense de Desenvolvimento Econômico e Social. IPARDES, Brasil*, 2009.

De Souza, L. *Programação genética e combinação de preditores para previsão de séries temporais*. Tese de Doutoramento, Universidade Federal do Paraná, 2006.

Eiben, A.; Hinterding, R.; Michalewicz, Z. Parameter control in evolutionary algorithms. *Evolutionary Computation, IEEE Transactions on*, v. 3, n. 2, p. 124 –141, 1999.

Farrar, C.; Duffey, T.; Doebling, S.; Nix, D. A statistical pattern recognition paradigm for vibration-based structural health monitoring. *Structural Health Monitoring*, p. 764–773, 1999.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996a.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; *et al.* Knowledge discovery and data mining: Towards a unifying framework. In: *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR*, 1996b, p. 82–88.

Forrest, S.; Nguyen, T.; Weimer, W.; Le Goues, C. A genetic programming approach to automated software repair. In: *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, ACM, 2009, p. 947–954.

Frawley, W.; Piatetsky-Shapiro, G.; Matheus, C. Knowledge Discovery in Databases: An Overview. *The American Association for Artificial Intelligence*, p. 57–70, 1992.

Freitas, A. On objective measures of rule surprisingness. *Principles of Data Mining and Knowledge Discovery*, p. 1–9, 1998.

FREITAS, A. A survey of evolutionary algorithms for data mining and knowledge discovery. *Advances in Evolutionary Computation*, p. 819–845, 2002a.

FREITAS, A. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Verlag, 2002b.

FURCY, D.; KOENIG, S. Limited discrepancy beam search. In: *International Joint Conference on Artificial Inteligente*, Citeseer, 2005, p. 125.

GEEM, Z. Optimal cost design of water distribution networks using harmony search. *Engineering Optimization*, v. 38, n. 3, p. 259–277, 2006.

GUIMARÃES, G.; LADEIRA, M.; GOMES, H. Disponibilização de aplicações inteligentes: Uma perspectiva de redução do custo médico da apnéia obstrutiva do sono não diagnosticada. *XXXIII Seminário Integrado de Software e Hardware*, 2006.

HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. Second. Morgan Kaufmann, 2006.

HAND, D. *Construction and assessment of classification rules*. John Wiley & Sons Inc, 1997.

HINTERDING, R.; MICHALEWICZ, Z.; EIBEN, A. Adaptation in evolutionary computation: a survey. In: *Evolutionary Computation, 1997., IEEE International Conference on*, 1997, p. 65 –69.

HOLLAND, J. H. *Adaptation in natural and artificial systems*. Cambridge, MA, USA: MIT Press, 1992.

HU, Y. A genetic programming approach to constructive induction. In: *Genetic Programming 1998: Proc. 3rd Annual Conf*, 1998, p. 146–151.

JIN, X.; REYNOLDS, R. Using knowledge-based evolutionary computation to solve nonlinear constraint optimization problems: a cultural algorithm approach. In: *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, 1999, p. 3 vol. (xxxvii+2348).

KIRKPATRICK, S.; GELATI JR, D.; VECCHI, M. Optimization by Simulated Annealing. *Science*, v. 220, p. 671–680, 1983.

KOH, H.; TAN, G. Data mining applications in healthcare. *Journal of Healthcare Information Management?Vol*, v. 19, n. 2, p. 65, 2005.

Koza, J. *Genetic programming: on the programming of computers by means of natural selection*. The MIT press, 1992.

Liu, B.; Hsu, W. Post-analysis of learned rules. In: *Proceedings of the National Conference on Artificial Intelligence*, Citeseer, 1996, p. 828–834.

Luke, S. Two fast tree-creation algorithms for genetic programming. *Evolutionary Computation, IEEE Transactions on*, v. 4, n. 3, p. 274 –283, 2000.

Maimon, O.; Rokach, L. *Data mining and knowledge discovery handbook*. Springer-Verlag New York Inc, 2005.

Mendel, J. Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE*, v. 83, n. 3, p. 345 –377, 1995.

Milley, A. Healthcare and data mining. *Health Management Technology*, v. 21, n. 8, p. 44–47, 2000.

Monard, M.; Baranauskas, J. Conceitos sobre aprendizado de máquina. *REZENDE, SO Sistemas inteligentes: fundamentos e aplicações. Barueri, SP: Manole*, p. 89–114, 2003.

Montana, D. J. Strongly typed genetic programming. *Evol. Comput.*, v. 3, n. 2, p. 199–230, 1995.

Neves, R. *Pré-processamento no processo de descoberta de conhecimento em banco de dados*. Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul - Instituto de Informática, 2003.

Obenshain, M. Application of data mining techniques to healthcare data. *Infection Control and Hospital Epidemiology*, p. 690–695, 2004.

Ortega, P.; Figueroa, C.; Ruz, G. A medical claim fraud/abuse detection system based on data mining: a case study in chile. *DMIN*, v. 6, p. 26–29, 2006.

Pappa, G. L. *Seleção de atributos utilizando algoritmos genéticos multiobjetivos*. Dissertação de Mestrado, Programa de Pós-Graduação em Informática Aplicada, Pontifícia Universidade Católica do Paraná, Curitiba - PR, 2002.

Pham, D.; Koç, E.; Lee, J.; Phrueksanant, J. Using the Bees Algorithm to schedule jobs for a machine. In: *Proc. of the 8th Int. Conf. and Exhibition on Laser Metrology, Machine Tool, CMM & Robotic Performance, LAMDAMAP*, 2007.

de Pinho, A. F.; Montevechi, J. A. B.; Marins, F. A. S. Análise da aplicação de projeto de experimentos nos parâmetros dos algoritmos genéticos. *Sistemas & Gestão 2*, v. 3, p. 319–331, 2007.

Quinlan, R. C5.0. Acesso em 15 dez. 2011, 2011.
Disponível em `http://www.rulequest.com/`

Rafalski, E. Using data mining/data repository methods to identify marketing opportunities in health care. *Journal of Consumer Marketing*, v. 19, n. 7, p. 607–613, 2002.

Reynolds, D.; Detroid, M. Cultural Algorithms: A Tutorial. In: *A tutorial given in the 2002 World Congress on*, 2002.

Reynolds, R. G.; Zanoni, E. Why cultural evolution can proceed faster than biological evolution. in: Proceedings of international symposium on simulating societies., p. 81–93. 1992.

Reynolds, R. An introduction to cultural algorithms. In: *Proceedings of the Third Annual Conference on Evolutionary Programming*, Rvier Edge, New Jersey: World Scientific Publishing, 1994, p. 131–139.

Reynolds, R. An overview of cultural algorithms. *Advances in Evolutionary Computation*, p. 367–378, 1999.

Reynolds, R.; Peng, B.; Brewster, J. Cultural swarms II: Virtual algorithm emergence. In: *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, 2003.

Reynolds, R.; PENG, B.; CHE, X. Knowledge Swarms: Generating Emergent Social Structure in Dynamic Environments. *Proceedings of the Agent 2005 Conference on Generative Social Processes, Models, and Mechanisms*, p. 747–770, 2001.

Reynolds, R.; Saleem, S. Cultural and Social Evolution in Dynamic Environments. *Thursday July 5, 2001*, p. 23, 2001.

Reynolds, R.; Zhu, S. Knowledge-based function optimization using fuzzy culturalalgorithms with evolutionary programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, v. 31, n. 1, p. 1–18, 2001.

Rezende, S. *Sistemas Inteligentes: fundamentos e aplicações*. Editora Manole Ltda, 2003.

Riaño, D. *Automatic construction of descriptive rules.* Tese de Doutoramento, Universitat Politècnica de Catalunya, Departament de LLenguatges I Sistemes Informàtics, 1997.

Romão, W.; Freitas, A. A.; Gimenes, I. M. Discovering interesting knowledge from a science & technology database with a genetic algorithm. *Applied Soft Computing Journal, Londres,*, v. 4, n. 2, p. 121–137, 2004.

Romão, W. *Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia.* Tese de Doutoramento, Universidade Federal de Santa Catarina, 2002.

Romão, W. Grupo de pesqusa em engenharia de algoritmo. Acesso em 15 dez. 2011, 2011.
Disponível em `www.din.uem.br/gpea/`

Shah-Hosseini, H. The intelligent water drops algorithm: a nature-inspired swarm-based optimization algorithm. *International Journal of Bio-Inspired Computation*, v. 1, n. 1, p. 71–79, 2009.

Siau, K. Health care informatics. *Information Technology in Biomedicine, IEEE Transactions on*, v. 7, n. 1, p. 1 –7, 2003.

Silberschatz, A.; Tuzhilin, A. What makes patterns interesting in knowledge discovery systems. *Knowledge and Data Engineering, IEEE Transactions on*, v. 8, n. 6, p. 970 –974, 1996.

Storn, R.; Price, K. Differential Evolution–A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization*, v. 11, p. 341–359, 1997.

Tan, P.-N.; Steinbach, M.; Kumar, V. *Introdução ao data mining: Mineração de dados.* Ciência Moderna, 2006.

Vianna, R.; Moro, C. Aplicação da descoberta de conhecimento em base de dados para a identificação de características relacionadas à mortalidade infantil. *Proceedings of International Association for Development of the Information Society. Lisboa: IADIS Press*, p. 678–80, 2005.

Witten, I. H.; Frank, E. *Data mining: Practical machine learning tools and techniques.* 2 ed. Morgan Kaufmann, 2005.

Xue, Z.; Guo, Y. Improved cultural algorithm based on genetic algorithm. In: *Integration Technology, 2007. ICIT '07. IEEE International Conference on*, 2007, p. 117 –122.

Zadeh, L. A. Fuzzy sets*. *Information and Control 8*, , n. 8, p. 338–353, 1965.

Apêndice

A

Descrição dos atributos dos dados de saúde

Atributo	Domínio	Descrição
BENEFICIARIO	[1..307491]	Código do Beneficiário
SEXO	F, M	Sexo do Beneficiário
IDADE	[0..105]	Idade do Beneficiário
ESTADO	SC, RS, PR, SP, RN, MG, RJ, DF, AL, MT, GO	Estado do Beneficiário
MUNICÍPIO	[ABDON BATISTA..ZORTEA]	Município do Beneficiário
TIPO-BENEF	TITULAR, DEPENDENTE	Tipo do Beneficiário
SADT	[0..285]	Serviço Auxiliar de Diagnóstico e Terapia
CONSULTA	[0..101]	Consulta Médica
INTERNAÇÃO	[0..63]	Internação Hospitalar
GRUPO-CID-I	[0..69]	Doenças Infecciosas e parasitárias
GRUPO-CID-II	[0..95]	Neoplasias

Atributo	Domínio	Descrição
GRUPO-CID-IV	[0..73]	Sangue, órgãos Hematopoiéticos e Transtornos Imunitários
GRUPO-CID-IIIA	[0..41]	Doenças Endócrinas
GRUPO-CID-IIIB	[0..49]	Doenças Nutricionais e metabólicas
GRUPO-CID-V	[0..272]	Doenças Mentais
GRUPO-CID-VIA	[0..175]	Sistema Nervoso
GRUPO-CID-VIB	[0..80]	Olhos e anexos, ouvido e apófise Mastóide
GRUPO-CID-VIIA	[0..144]	Circulatório, menos Veias e Linfáticos
GRUPO-CID-VIIB	[0..39]	Veias e Linfáticos
GRUPO-CID-VIIIA	[0..74]	Infecções Respiratórias Agudas
GRUPO-CID-VIIIB	[0..64]	Outras Doenças Respiratórias
GRUPO-CID-VIIIC	[0..46]	Doenças crônicas das vias aéreas Inferiores
GRUPO-CID-IXA	[0..138]	Doenças do esôfago, estômago e duodeno
GRUPO-CID-IXB	[0..40]	Doenças do apêndice
GRUPO-CID-IXC	[0..40]	Hérnias da cavidade abdominal
GRUPO-CID-IXD	[0..44]	Doenças dos intestinos e peritôneo
GRUPO-CID-IXE	[0..70]	Outras doenças do aparelho digestivo
GRUPO-CID-IXF	[0..24]	Doenças do fígado
GRUPO-CID-IXG	[0..31]	Vesícula biliar, vias biliares e pâncreas
GRUPO-CID-XII	[0..48]	Pele e tecido subcutâneo
GRUPO-CID-XIII	[0..115]	Osteomuscular e tecido conjuntivo
GRUPO-CID-XA	[0..56]	Doenças Urinárias
GRUPO-CID-XB	[0..85]	Doenças genitais masculinas
GRUPO-CID-XC	[0..117]	Doenças genitais femininas
GRUPO-CID-XD	[0..75]	Insuficiência renal
GRUPO-CID-XV	[0..4]	Afecções originadas no período perinatal
GRUPO-CID-XI	[0..22]	Gravidez, parto e puerpério

Atributo	**Domínio**	**Descrição**
GRUPO-CID-XIV	[0..55]	Malformações Congênitas
GRUPO-CID-XVI	[0..93]	Sintomas, sinais e afecções mal definidas
GRUPO-CID-XVII	[0..82]	Lesões, envenenamentos e causas externas
GRUPO-CID-XXI	[0..66]	Contato com os serviços de saúde
ESPECIALIDADE-01	[0..5]	Acupuntura
ESPECIALIDADE-02	[0..6]	Administração
ESPECIALIDADE-03	[0..4]	ALERGIA E I
ESPECIALIDADE-04	[0..16]	Análises Clínicas
ESPECIALIDADE-05	[0..8]	Anatomia Patológica
ESPECIALIDADE-06	[0..6]	Anestesiologia
ESPECIALIDADE-07	[0..7]	Angiologia
ESPECIALIDADE-08	[0..9]	Cancerologia
ESPECIALIDADE-09	[0..16]	Cardiologia
ESPECIALIDADE-10	[0..4]	CIRURGIA CA
ESPECIALIDADE-11	[0..6]	CIRURGIA DA
ESPECIALIDADE-12	[0..3]	CIRURGIA DO
ESPECIALIDADE-13	[0..8]	Cirurgia Geral
ESPECIALIDADE-14	[0..5]	CIRURGIA PE
ESPECIALIDADE-15	[0..6]	CIRURGIA PL
ESPECIALIDADE-16	[0..4]	CIRURGIA TO
ESPECIALIDADE-17	[0..4]	CIRURGIA VA
ESPECIALIDADE-18	[0..56]	Clínica Médica
ESPECIALIDADE-19	[0..5]	Coloproctologia
ESPECIALIDADE-20	[0..5]	Consulta Especializada
ESPECIALIDADE-21	[0..2]	Consulta Especializada
ESPECIALIDADE-22	[0..3]	Consulta Especializada
ESPECIALIDADE-23	[0..2]	Densitometria
ESPECIALIDADE-24	[0..7]	Dermatologia
ESPECIALIDADE-25	[0..2]	Ecocardiograma
ESPECIALIDADE-26	[0..2]	Eletrocardiograma
ESPECIALIDADE-27	[0..4]	Endocrinologia
ESPECIALIDADE-28	[0..4]	Endoscopia

Atributo	Domínio	Descrição
ESPECIALIDADE-29	[0..12]	Fisioterapia
ESPECIALIDADE-30	[0..2]	Fonoaudiologia
ESPECIALIDADE-31	[0..5]	Gastroenterologia
ESPECIALIDADE-32	[0..3]	Geriatria
ESPECIALIDADE-33	[0..12]	Ginecologia
ESPECIALIDADE-34	[0..7]	Hematologia
ESPECIALIDADE-35	[0..3]	Hemodialise
ESPECIALIDADE-36	[0..3]	Homeopatia
ESPECIALIDADE-37	[0..1]	Hospital
ESPECIALIDADE-38	[0..4]	Infectologia
ESPECIALIDADE-39	[0..23]	Internações
ESPECIALIDADE-40	[0..4]	Mastologia
ESPECIALIDADE-41	[0..17]	MED. FÂSICA
ESPECIALIDADE-42	[0..3]	MEDICINA DE
ESPECIALIDADE-43	[0..4]	MEDICINA DO
ESPECIALIDADE-44	[0..3]	MEDICINA IN
ESPECIALIDADE-45	[0..1]	MEDICINA LE
ESPECIALIDADE-46	[0..1]	MEDICINA NU
ESPECIALIDADE-47	[0..12]	Medicina Preventiva
ESPECIALIDADE-48	[0..9]	Nefrologia
ESPECIALIDADE-49	[0..4]	Neurocirurgia
ESPECIALIDADE-50	[0..5]	Neurologia
ESPECIALIDADE-51	[0..3]	Nutricionismo
ESPECIALIDADE-52	[0..3]	Nutrologia
ESPECIALIDADE-53	[0..9]	Oftalmologia
ESPECIALIDADE-54	[0..11]	Ortopedia
ESPECIALIDADE-55	[0..9]	Otorrinolaringologia
ESPECIALIDADE-56	[0..6]	Patologia
ESPECIALIDADE-57	[0..11]	Patologia Clínica
ESPECIALIDADE-58	[0..13]	Pediatria
ESPECIALIDADE-59	[0..6]	Pneumologia
ESPECIALIDADE-60	[0..2]	Pronto Socorro
ESPECIALIDADE-61	[0..3]	Psicoterapia
ESPECIALIDADE-62	[0..6]	Psiquiatria

Atributo	**Domínio**	**Descrição**
ESPECIALIDADE-63	[0..5]	Radiodiagnostico
ESPECIALIDADE-64	[0..9]	Radiologia
ESPECIALIDADE-65	[0..11]	Radioterapia
ESPECIALIDADE-66	[0..4]	Reumatologia
ESPECIALIDADE-67	[0..3]	Tomografia
ESPECIALIDADE-68	[0..2]	Ultrassonografia
ESPECIALIDADE-69	[0..9]	Urologia
GRUPO-EVENTO-01	[0..178]	Procedimentos - DM - Diabetes Mellitus
GRUPO-EVENTO-02	[0..121]	Procedimentos - HAS - Hipertensão Arterial
GRUPO-EVENTO-03	[0..0]	Procedimentos - Todos
GRUPO-EVENTO-04	[0..19]	Procedimentos Sentinelas - DM - Diabetes Mellitus
GRUPO-EVENTO-05	[0..1]	Procedimentos Sentinelas - HAS - Hipertensão
GRUPO-EVENTO-06	[0..85]	Procedimentos Sentinelas - Câncer de colo de Útero
GRUPO-EVENTO-07	[0..45]	Procedimentos Sentinelas - Câncer de Mama
GRUPO-EVENTO-08	[0..0]	Medicamentos - DM - Diabetes Mellitus
GRUPO-EVENTO-09	[0..45]	Procedimentos Câncer de Mama
GRUPO-EVENTO-10	[0..85]	Procedimentos Câncer de Colo de Útero
GRUPO-EVENTO-11	[0..2]	GRUPOS PROC/MED2 GESIANE LEAL

Apêndice

B

Tabela de frequência dos atributos de especialidade

Especialidade	**Frequência**	**Freq. Sem Rep.**
ESPECIALIDADE-01	335	247
ESPECIALIDADE-02	167	146
ESPECIALIDADE-03	662	493
ESPECIALIDADE-04	7942	4833
ESPECIALIDADE-05	1169	247
ESPECIALIDADE-06	1900	1607
ESPECIALIDADE-07	1153	879
ESPECIALIDADE-08	772	499
ESPECIALIDADE-09	14040	8275
ESPECIALIDADE-10	136	88
ESPECIALIDADE-11	236	176
ESPECIALIDADE-12	310	282
ESPECIALIDADE-13	4779	3759
ESPECIALIDADE-14	365	281
ESPECIALIDADE-15	356	297
ESPECIALIDADE-16	223	157
ESPECIALIDADE-17	251	213

Especialidade	**Frequência**	**Freq. Sem Rep.**
ESPECIALIDADE-18	132557	51106
ESPECIALIDADE-19	605	480
ESPECIALIDADE-20	597	498
ESPECIALIDADE-21	94	83
ESPECIALIDADE-22	99	78
ESPECIALIDADE-23	80	77
ESPECIALIDADE-24	5116	3926
ESPECIALIDADE-25	35	33
ESPECIALIDADE-26	33	30
ESPECIALIDADE-27	2915	2564
ESPECIALIDADE-28	283	239
ESPECIALIDADE-29	897	589
ESPECIALIDADE-30	25	22
ESPECIALIDADE-31	2620	2086
ESPECIALIDADE-32	769	689
ESPECIALIDADE-33	12519	8886
ESPECIALIDADE-34	523	415
ESPECIALIDADE-35	4	2
ESPECIALIDADE-36	776	632
ESPECIALIDADE-37	19	19
ESPECIALIDADE-38	156	127
ESPECIALIDADE-39	4943	3095
ESPECIALIDADE-40	227	183
ESPECIALIDADE-41	798	458
ESPECIALIDADE-42	221	189
ESPECIALIDADE-43	317	268
ESPECIALIDADE-44	82	73
ESPECIALIDADE-45	1	1
ESPECIALIDADE-46	163	163
ESPECIALIDADE-47	6082	3805
ESPECIALIDADE-48	703	585
ESPECIALIDADE-49	662	577
ESPECIALIDADE-50	2310	1905
ESPECIALIDADE-51	40	27

Especialidade	**Frequência**	**Freq. Sem Rep.**
ESPECIALIDADE-52	281	217
ESPECIALIDADE-53	14901	9390
ESPECIALIDADE-54	10930	7948
ESPECIALIDADE-55	6900	4872
ESPECIALIDADE-56	4942	4321
ESPECIALIDADE-57	18217	12573
ESPECIALIDADE-58	14548	9583
ESPECIALIDADE-59	2462	1762
ESPECIALIDADE-60	22	16
ESPECIALIDADE-61	19	12
ESPECIALIDADE-62	4259	3040
ESPECIALIDADE-63	1676	1337
ESPECIALIDADE-64	9115	6689
ESPECIALIDADE-65	414	271
ESPECIALIDADE-66	1142	980
ESPECIALIDADE-67	174	156
ESPECIALIDADE-68	117	100
ESPECIALIDADE-69	3537	2849

Apêndice

C

Distribuição de classes dos dados de saúde

Atributo meta	**Descrição**	**SIM(%)**	**NÃO(%)**
GRUPO-CID-I	Doenças Infecciosas e parasitárias	2,86	97,14
GRUPO-CID-II	Neoplasias	4,07	95,93
GRUPO-CID-IIIA	Doenças Endócrinas	2,26	97,74
GRUPO-CID-IIIB	Doenças Nutricionais e metabólicas	1,85	98,16
GRUPO-CID-V	Doenças Mentais	2,42	97,58
GRUPO-CID-VIA	Sistema Nervoso	1,87	98,13
GRUPO-CID-VIB	Olhos e anexos, ouvido e apófise Mastóide	4,70	95,30
GRUPO-CID-VIIA	Circulatório, menos Veias e Linfáticos	10,12	89,88
GRUPO-CID-VIIIA	Infecções Respiratórias Agudas	4,67	95,33
GRUPO-CID-VIIIB	Outras Doenças Respiratórias	1,73	98,28
GRUPO-CID-IXA	Doenças do esôfago, estômago e duodeno	1,81	98,19
GRUPO-CID-XII	Pele e tecido subcutâneo	2,71	97,29
GRUPO-CID-XIII	Osteomuscular e tecido conjuntivo	5,47	94,53
GRUPO-CID-XA	Doenças Urinárias	3,00	97,00
GRUPO-CID-XB	Doenças genitais masculinas	1,73	98,27
GRUPO-CID-XC	Doenças genitais femininas	5,89	94,11
GRUPO-CID-XVI	Sintomas, sinais e afecções mal definidas	10,55	89,45
GRUPO-CID-XXI	Contato com os serviços de saúde	17,75	82,25

Atributo meta	**Descrição**	**SIM(%)**	**NÃO(%)**
ESPECIALIDADE-04	Análises Clínicas	3,51	96,49
ESPECIALIDADE-05	Anatomia Patológica	0,52	99,49
ESPECIALIDADE-06	Anestesiologia	1,17	98,83
ESPECIALIDADE-07	Angiologia	0,64	99,36
ESPECIALIDADE-09	Cardiologia	6,02	93,98
ESPECIALIDADE-13	Cirurgia Geral	2,73	97,27
ESPECIALIDADE-18	Clínica Médica	37,13	62,87
ESPECIALIDADE-24	Dermatologia	2,86	97,14
ESPECIALIDADE-27	Endocrinologia	1,87	98,13
ESPECIALIDADE-31	Gastroenterologia	1,52	98,48
ESPECIALIDADE-33	Ginecologia	6,46	93,54
ESPECIALIDADE-39	Internações	2,24	97,76
ESPECIALIDADE-47	Medicina Preventiva	2,77	97,23
ESPECIALIDADE-50	Neurologia	1,38	98,62
ESPECIALIDADE-53	Oftalmologia	6,82	93,18
ESPECIALIDADE-54	Ortopedia	5,78	94,22
ESPECIALIDADE-55	Otorrinolaringologia	3,54	96,46
ESPECIALIDADE-56	Patologia	3,14	96,86
ESPECIALIDADE-57	Patologia Clínica	9,14	90,86
ESPECIALIDADE-58	Pediatria	6,69	93,31
ESPECIALIDADE-59	Pneumologia	1,28	98,72
ESPECIALIDADE-62	Psiquiatria	2,22	97,79
ESPECIALIDADE-63	Radiodiagnostico	0,97	99,03
ESPECIALIDADE-64	Radiologia	4,85	95,15
ESPECIALIDADE-66	Reumatologia	0,71	99,29
ESPECIALIDADE-69	Urologia	2,07	97,93
GRUPO-EVENTO-01	Procedimentos- Diabetes Mellitus	0,67	99,33
GRUPO-EVENTO-04	Procedimentos Sentinelas - Diabetes Mellitus	0,44	99,56
GRUPO-EVENTO-09	Procedimentos Câncer de Mama	3,62	96,39