UNIVERSIDADE FEDERAL DE ALFENAS
INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Juliana Oliveira Ferreira

DIAGNOSTICAR MALIGNIDADE DO CÂNCER DE BOCA COM AUXÍLIO DA PROGRAMAÇÃO GENÉTICA

Alfenas, 01 de Julho de 2010.

UNIVERSIDADE FEDERAL DE ALFENAS
INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DIAGNOSTICAR MALIGNIDADE DO CÂNCER DE BOCA COM AUXÍLIO DA PROGRAMAÇÃO GENÉTICA

Juliana Oliveira Ferreira

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Alfenas como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Orientador: Prof. Humberto César Brandão de Oliveira
Co-orientador: Prof. Dr. Alessandro Antônio Costa Pereira

Alfenas, 01 de Julho de 2010.

Juliana Oliveira Ferreira

DIAGNOSTICAR MALIGNIDADE DO CÂNCER DE BOCA COM AUXÍLIO DA PROGRAMAÇÃO GENÉTICA

A Banca examinadora abaixo-assinada aprova a monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação pela Universidade Federal de Alfenas.

Prof. Dr. Eric Batista Ferreira

Universidade Federal de Alfenas

Prof. Dr. Ricardo Menezes Salgado

Universidade Federal de Alfenas

Prof. Humberto César Brandão de Oliveira (Orientador)

Universidade Federal de Alfenas

Alfenas, 01 de julho de 2010.

Dedico este trabalho aos meus Pais por estarem presentes em todos os momentos de minha vida e por não terem medido esforços para a minha educação.

AGRADECIMENTO

Agradeço primeiramente a Deus por permitir a concretização deste trabalho e pela força dada nos momentos mais difíceis desta caminhada.

Aos meus queridos pais, Arildo Barbosa Ferreira e Cássia Junqueira Oliveira Ferreira, por encherem meu coração de amor e principalmente por me fazerem sentir especial.

As minhas irmãs, Adriana Oliveira Ferreira e Patrícia Oliveira Ferreira, por serem minhas melhores amigas.

Ao meu namorado, Saulo Araújo Naves, por simplesmente tornar meus dias mais alegres e especiais.

A minha madrinha, Neiva Barbosa Ferreira, pelo apoio e suas palavras de carinho.

Ao meu orientador, Prof. Humberto César Brandão de Oliveira, pelos conhecimentos transmitidos, amizade e acima de tudo pela oportunidade e paciência ao longo deste trabalho.

A Universidade Federal de Alfenas (UNIFAL) e ao Laboratório de Inteligência Computacional (LInC), que foram os alicerces para esta realização.

A FAPEMIG pelo apoio financeiro.

A Profª. Dr. Melise Maria Veiga de Paula e ao Prof. Dr. Alessandro Antônio Costa Pereira pela orientação e pelo conhecimento proporcionado.

Ao Prof. Dr. Luiz Alberto Beijo e ao Prof. Luiz Eduardo da Silva pela ajuda e esclarecimento.

A todos os mestres pela dedicação e pelos ensinamentos.

Aos meus colegas, pela amizade e companheirismo. Em especial ao Edgar, Max e Lucas Schmoeller, pessoas as quais estimo muito.

A todos aqueles que de alguma forma ajudaram na realização deste trabalho. Muito Obrigada,

Juliana Oliveira Ferreira.

"Existe uma coisa que uma longa existência me ensinou: toda a nossa ciência, comparada à realidade, é primitiva e inocente; e portanto, é o que temos de mais valioso"

Albert Einstein

RESUMO

Quando um paciente é diagnosticado com uma lesão cancerígena na boca, este deve ser submetido a alguns procedimentos, na tentativa de quantificar a graduação desta lesão e a sua extensão e disseminação aparente. O termo graduação refere-se ao estabelecimento de uma estimativa da agressividade ou nível de malignidade da lesão e geralmente é feita com base na diferenciação citológica das células tumorais. Entretanto, existem várias abordagens para esta classificação e não há, até o presente momento, consenso sobre a classificação histopatológica que possua maior poder preditivo. Isto dificulta a realização de exames por profissionais menos experientes e torna possível que uma mesma lesão seja diagnosticada de forma diferente por mais de um especialista.

Com base nestas informações, este trabalho busca desenvolver um sistema computacional inteligente que seja capaz de realizar esta classificação, auxiliando na escolha da terapia mais adequada e na estimativa de sobrevida dos pacientes. O sistema corresponde a uma adaptação da programação genética e consiste em localizar expressões matemáticas, que ao serem processadas com características numéricas das células da lesão, possibilitem a classificação desta. As diferentes fórmulas aritméticas nesta metodologia são definidas em uma especifica linguagem livre de contexto, e a representação computacional é feita através de árvores sintáticas.

Testes foram realizados e através dos resultados pode-se concluir que o método proposto pode ser uma boa ferramenta para auxiliar os patologistas no diagnostico do câncer. A técnica também foi capaz de preservar o conhecimento de profissionais experientes e selecionar as características importantes para realizar a classificação, diminuindo assim, o custo e o tempo gastos na coleta de dados de futuras classificações.

Palavras-Chave: Classificação de Padrões, Programação Genética, Células cancerígenas.

ABSTRACT

When a patient is diagnosed with a cancerous lesion in the mouth, it must undergo some procedures in an attempt to quantify the degree of this lesion and its extent and apparent spread. This graduation is to establish an estimate of their level of aggressiveness or malignancy and it's usually made based on the cytological differentiation of tumor cells. However, there are several approaches to this classification and there is so far no consensus on the histopathologic classification that has greater predictive power. This complicates the implementation of tests for less experienced professionals and enables the same lesion be diagnosed differently by more than one specialist.

Based on this information, this paper aims to develop a system that is capable of performing this classification, helping to choose the most appropriate therapy and survival rate of patients. The system represents an adaptation of genetic programming and look for mathematical expressions, which are processed with the numerical characteristics of cells in the lesion, allowing the classification of this one. The several arithmetic formulas in this methodology are defined in a specific free context language, and computational representation is made by parse tree.

Tests were conducted and by the results we can conclude that the proposed method can be a good tool to assist pathologists in diagnosing cancer. The technique was also able to preserve the knowledge of experienced professionals and to select the important features to perform the classification, thereby reducing the cost and time spent collecting data for future classifications.

Keywords: Pattern Classification, Genetic Programming, Cancer cells.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

AGs	Algoritmos Genéticos
CCE	Carcinomas de células escamosas
CP	Classificação de Padrões
IA	Inteligência Artificial
INCA	Instituto Nacional do Câncer
MLP	Multilayer Perceptron (Rede Neural Perceptron Multicamada)
OMS	Organização Mundial de Saúde
PG	Programação Genética
RNA	Redes Neurais Artificiais
SGBD	Sistema de gerenciamento de Banco de Dados
SQE	Soma de quadrados dos erros (*Sum Square Error*)
UNIFAL	Universidade Federal de Alfenas

SUMÁRIO

1
Introdução

Este capítulo tem por finalidade apresentar as principais motivações, os objetivos e as contribuições do presente trabalho.

Esta pesquisa apresenta um sistema computacional inteligente utilizado para classificar o grau de desenvolvimento de lesões cancerígenas da cavidade oral. Esse sistema foi desenvolvido como uma ferramenta de apoio decisório da escolha terapêutica e pode contribuir com o patologista para uma classificação mais objetiva e uniforme dos tumores analisados, acelerando o trabalho desse profissional e estimando a sobrevida do paciente.

A técnica utilizada para realizar a classificação corresponde a uma adaptação da programação genética (PG) e consiste em localizar expressões matemáticas, que ao serem processadas com características numéricas da lesão em análise, possibilitam a classificação.

O sistema funciona como uma forma de preservar, aproveitar e organizar o talento e a experiência de especialistas, sendo uma de suas principais vantagens a capacidade de armazenar o conhecimento de mais de um profissional. Essa vantagem pode diminuir consideravelmente o custo e o tempo gastos em casos que o patologista considere necessário a avaliação da lesão por mais de um especialista.

Como segundo plano, a técnica procura determinar quais características são realmente necessárias para a realização desta classificação, ou seja, se um especialista determina que quantidade de mitoses, queratinização e pleomorfismo são características importantes para determinar o grau de malignidade de uma lesão, a técnica pode mostrar, se isso for possível, que com apenas duas das características selecionadas consegue-se realizar a classificação com alta taxa de precisão.

Para realização dos experimentos, este trabalho utilizou uma base de dados desenvolvida pela disciplina de Patologia da Universidade Federal de Alfenas e três bases de dados reais e bem exploradas pela literatura, encontradas no

repositório Proben1. As bases de dados deste repositório que foram utilizadas pela pesquisa são:

- Câncer de mama: Determina se um nódulo na mama é maligno ou benigno;
- Diabetes: Determina se um paciente é diabético ou não;
- Destino de cavalos com cólicas: prevê o destino de cavalos com cólicas indicando se o cavalo vai sobreviver, morrer, ou se deve ser sacrificado.

Através desta diversificação de bases de dados é possível avaliar a adaptabilidade da técnica a outros tipos de classificação.

1.1 Justificativa e Motivação

O diagnóstico laboratorial de tumores na extremidade maligno-benigno, na maioria das vezes não é algo difícil. Porém, a quantificação da provável agressividade clínica e a extensão e disseminação aparente de lesões malignas são necessárias para a realização de um prognóstico mais preciso e para a comparação de resultados finais de vários protocolos de tratamento (Abbas *et al.*, 2008).

A quantificação da provável agressividade clínica, conhecida também como graduação, é uma tentativa de estabelecer o nível de malignidade com base na diferenciação citológica das células tumorais e no número de mitoses dentro do tumor, auxiliando na escolha do tratamento e estimando a sobrevida do paciente (Abbas *et al.*, 2008). Entretanto, existem várias formas de realizar esta classificação, e até o presente momento não há um consenso sobre qual delas possui maior poder preditivo, tornando o exame subjetivo e dificultando a realização deste por profissionais menos experientes (Lourenço *et al.*, 2007).

Sendo assim, esta pesquisa tem como foco principal desenvolver um Classificador de Células malignas com o objetivo de facilitar e tornar mais confiável a realização desses exames, procurando aumentar a taxa de sucesso dos tratamentos e proporcionar melhor bem estar a pacientes que não precisão de terapias mais agressivas.

1.2 Problematização

As classificações histopatológicas para carcinomas celulares da boca, foco desta pesquisa, surgiram na tentativa de explicar o comportamento biológico discrepante dos tumores com características clínicas semelhantes. Essa classificação facilita a escolha da terapia mais adequada e estima a sobrevida dos pacientes. Entretanto, existem diversos sistemas de graduação histopatológica e, até o presente momento, não há na literatura consenso sobre esta classificação (Abbas *et al.*, 2008).

Os exames realizados para determinar o estágio histopatológico destes tumores são exames subjetivos e variam com a experiência do patologista. Além disso, é pratica comum o envio de amostras da lesão para vários patologistas para que estes entrem em consenso, quando o especialista não esta muito seguro do diagnóstico. Desta forma, como é possível tornar mais fácil e confiável a realização destes exames?

1.3 Objetivos

1.3.1 Gerais

Este trabalho objetivou construir um sistema computacional inteligente capaz de diferenciar células cancerígenas com satisfatório grau de precisão. A idéia principal é diferenciar lesões localizadas na região da boca e pescoço de acordo com o grau de desenvolvimento da lesão. Como segundo plano, a técnica procura determinar quais características são realmente necessárias para a realização desta classificação, diminuindo assim o custo na coleta de futuras classificações.

1.3.2 Específicos

Outros objetivos foram visados pelo presente trabalho, entre eles:

- Implantar o sistema no Departamento de Patologia da Universidade Federal de Alfenas – UNIFAL.

- Preservar o conhecimento de vários especialistas através das bases de dados elaboradas.
- Divulgar os conhecimentos obtidos e resultados alcançados pelo projeto.

1.4 Organização da Monografia

Este trabalho encontra-se dividido da seguinte forma:

O Capítulo 2 apresenta alguns aspectos teóricos relacionados a classificação de células cancerígenas e utilização de sistemas inteligentes para classificação de padrões.

O Capítulo 3 tem como objetivo apresentar alguns métodos conhecidos para a classificação oncológica de células cancerígenas e algumas técnicas já conhecidas na literatura, propostas para resolver o problema de classificação de padrões.

O Capítulo 4 apresenta o funcionamento básico de um sistema de classificação e descreve a nova técnica proposta pelo presente trabalho. Como o sistema proposto consiste em uma adaptação da programação genética, este capítulo apresenta também uma descrição desta técnica.

O Capítulo 5 apresenta o sistema de informação desenvolvido para determinar o grau de malignidade de casos de câncer de boca. Será apresentada a arquitetura básica do sistema, além de algumas interfaces gráficas desenvolvidas para este.

O Capítulo 6 é o responsável por exibir os resultados alcançados pela pesquisa, sendo apresentado também as bases de dados e as configurações utilizadas pelo sistema durante os experimentos e alguns gráficos e tabelas para analisarmos os resultados.

As conclusões gerais do projeto, vantagens e trabalhos futuros estão apresentados no capítulo 7.

2
Classificação de Células Cancerígenas

Neste capítulo serão descritos alguns aspectos teóricos relacionados a esta área que fundamentam a proposta apresentada neste trabalho. Na Seção 2.1 será apresentada a importância da classificação de células cancerígenas, a Seção 2.2, apresenta conceitos básicos sobre exames realizados para o diagnostico do câncer, e na Seção 2.3, será apresentada a descrição de sistemas classificadores e algumas áreas do conhecimento onde podem ser utilizadas.

2.1 Importância

O câncer bucal consiste em lesões que acometem a boca e parte da garganta. Pode se desenvolver nos lábios, gengivas, mucosa jugal (bochechas), palato duro (céu da boca), língua (principalmente as margens), assoalho (região embaixo da língua) e amígdalas (INCA, 2010a). O número de casos vem apresentando um aumento significativo em todo o mundo, configurando-se atualmente em um dos mais importantes problemas de saúde pública mundial (INCA, 2009). Na Figura 1, observa-se a taxa de mortalidade proporcional causada pelo câncer de boca, no Brasil, entre 1979 e 2007. Estes dados foram obtidos através do Instituto Nacional do Câncer (2010b).

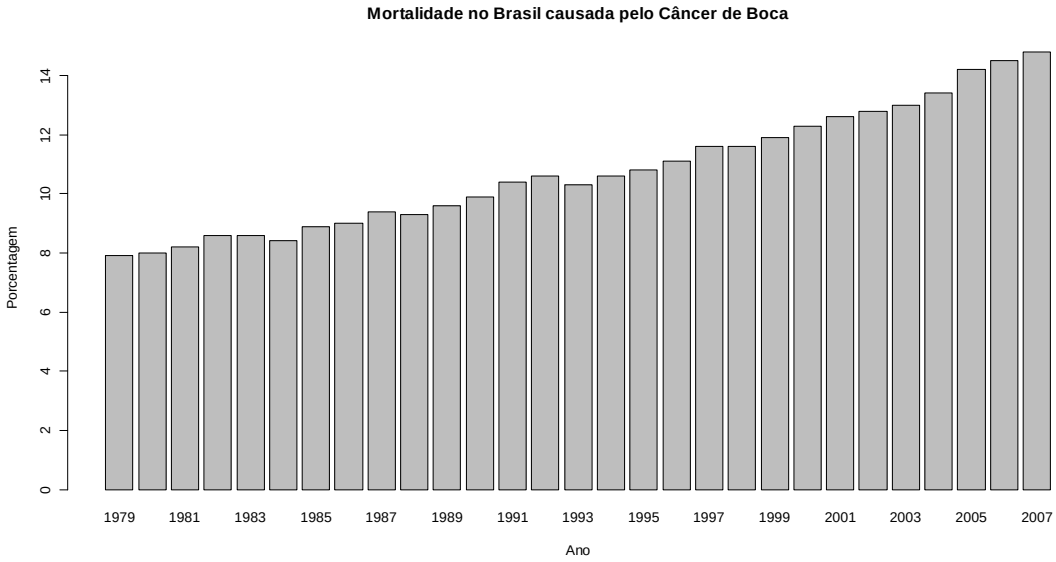

Figura 1. Mortalidade por câncer, no Brasil, entre 1997 e 2007.

Segundo o INCA (2010a), no Brasil, a estimativa de novos casos para 2010 encontra-se em 14.120, sendo 10.330 homens e 3.790 mulheres, e o número de mortes, estimados em 6.214, sendo 4.898 homens e 1.316 mulheres. Na Figura 2 e Figura 3, pode-se observar que o câncer bucal apresenta uma distribuição geográfica variável. A Figura 2 apresenta a taxa de incidência de câncer de boca para 2010, estimada para cada 100.000 homens, e a Figura 3 para cada 100.000 mulheres.

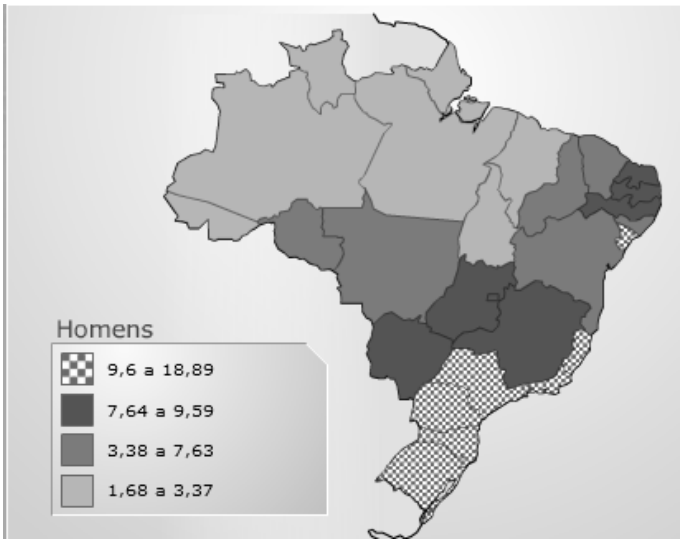

Figura 2. Taxa de incidência do câncer de boca, por 100.000 homens, estimados para o ano de 2010 (INCA, 2009).

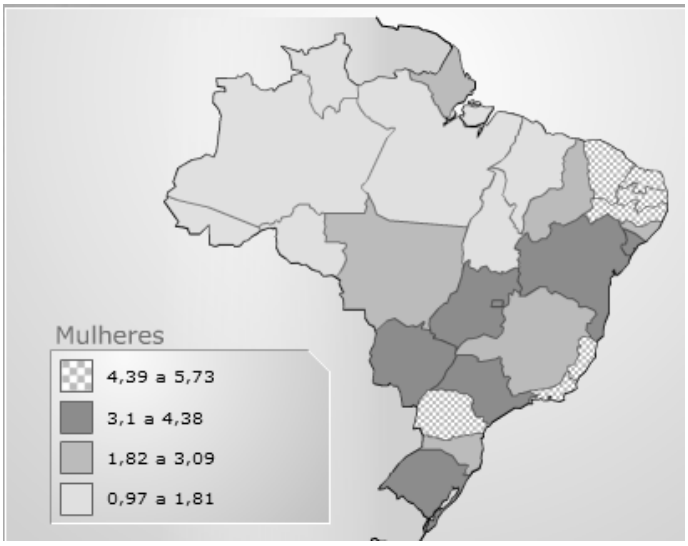

Figura 3. Taxa de incidência do câncer de boca, por 100.000 mulheres, estimados para o ano de 2010 (INCA, 2009).

Levando em consideração as altas taxas informadas anteriormente, é indispensável a definição de um diagnóstico seguro para que o paciente possa ser submetido ao tratamento mais adequado, estimando assim, a sua sobrevida. Essa definição obrigatoriamente tem sido feita através de métodos que determinam o grau de malignidade e o estadiamento clínico da lesão. Embora ambas estejam correlacionadas, são duas maneiras independentes de avaliar o comportamento de um tumor.

O estadiamento, ou extensão e disseminação aparente da lesão, é geralmente avaliado através de exames clínicos e radiológicos e baseia-se na extensão local da lesão, sua disseminação para os linfonodos regionais e presença ou ausência de metástase. Tem como principal função caracterizar os tumores e auxiliar na escolha terapêutica (Lourenço *et al.*, 2007; Abbas *et al.*, 2008).

Já o grau de malignidade, também conhecido como agressividade clínica, é atribuído por um patologista. Baseia-se em exames histopatológicos, onde são avaliadas, através de exame microscópio, células coletadas das lesões. Esta classificação histopatológica procura estabelecer uma estimativa da sua agressividade e é o foco desta pesquisa. Suas principais importâncias estão na tentativa de explicar o comportamento biológico discrepante, prover fatores prognósticos suplementares e auxiliar especialistas na escolha terapêutica mais

adequada estimando a sobrevida do paciente (Lourenço *et al.*, 2007; Abbas *et al.*, 2008; Costa et al., 2002).

2.2 Método de classificação existente (Biópsia)

O câncer de boca é uma denominação que inclui tanto lesões localizadas nos lábios como as localizadas no interior da boca (mucosa bucal, gengivas, palato duro, língua oral e assoalho da boca) (INCA, 2006b). Quando encontrado em um paciente, este deve submeter-se a alguns procedimentos com a finalidade de quantificar a sua provável agressividade. Para esta classificação é necessário a realização de um exame histopatológico, denominado biópsia, na qual é retirado material para determinar as características do tumor. Depois de coletado, esse material deve ser enviado ao laboratório de patologia, onde passa por um processo específico para a preparação de lâminas que serão observadas através de um microscópio (Ferreira, 2010).

Existem várias abordagens para a retirada deste material, incluindo:

- Biópsia Total (Excisional): indicada para tumores pequenos e superficiais e consiste no procedimento no qual toda a lesão é retirada (Ferreira, 2010; Lourenço *et al.*, 2007);
- Biópsia Parcial (Incisional): remove-se um fragmento da lesão através de incisão cirúrgica, indicado para tumores maiores e profundos (Ferreira, 2010; Lourenço *et al.*, 2007);
- Biópsia por agulha: o material é coletado através de punção, sem cirurgia. Inclui tanto as punções aspirativas com agulha fina, na qual se obtém material para avaliação do aspecto das células, como as biópsias por agulhas grossas, que consistem em obter fragmentos de tecido (Ferreira, 2010; Lourenço *et al.*, 2007);
- Biópsia por esfregaços citológicos – o material é obtido pela raspagem na superfície da lesão, sendo considerado um método não-invasivo, pois o material é facilmente coletado (Lourenço *et al.*, 2007).

A forma pela qual o material será coletado deve ser escolhida pelo especialista levando-se em consideração os aspectos clínicos da lesão. Em alguns casos essa coleta deve ser orientada por exames de imagem, a fim de que o local de retirada seja mais precisamente identificado.

O resultado da biópsia pode levar alguns dias, pois como apresentado anteriormente, após a coleta, o material deve ser submetido a técnicas especiais para a preparação da lâmina. O próximo passo é avaliar essas lâminas através de microscópios, sendo o especialista responsável pala classificação da lesão.

O critério de classificação utilizado varia conforme cada forma de neoplasia (Abbas *et al*., 2008). No caso dos carcinomas de células escamosas (CCE) da boca (Lourenço *et al.*, 2007), foco desta pesquisa, podem-se citar:

- Classificação Descritiva;
- Sistema de Graduação Multifatorial;
- Sistema de classificação das Margens Invasivas;
- Avaliação Histopatológicas de Risco.
- Gradação Histopatológica da Organização Mundial de Saúde (OMS).

A forma de classificação utilizada varia entre os especialistas e, até o presente momento, não há na literatura consenso sobre a classificação histopatológica que possua maior poder preditivo (Lourenço *et al.*, 2007). Mais detalhes sobre as formas de classificação podem ser observadas na Seção 3.1.

2.3 Sistemas Inteligentes para Classificação de Padrões

As pesquisas da área de Inteligência Artificial (IA) consistem em elaborar sistemas inteligentes, que possibilitem aos computadores realizar funções que são desempenhadas pelos seres humanos utilizando conhecimento e raciocínio (Rezende *et al.*, 2003, pág. 3). Entre estas funções encontra-se a de classificar objetos de acordo com algumas características disponíveis. Por exemplo, ele é capaz de avaliar características de uma pessoa (tamanho, peso, comprimento do cabelo, tom

de voz, entre outras) e determinar se esta é do sexo masculino ou do sexo feminino. Outro exemplo é o indicado na Seção 2.2, onde analisando características das células de uma lesão maligna, patologistas são capazes de classificá-las.

Esses sistemas são denominados classificadores de padrões e consistem basicamente em encontrar propriedades comuns entre um conjunto de instâncias em um banco de dados, permitindo assim a classificação de novas entradas (Eiben & Smith, 2003; Lin, 2007; Tsakonas, 2006).

Os sistemas classificadores podem ser utilizados em diversas áreas do conhecimento como sistemas de apoio à decisão. Entre essas áreas, pode-se citar como exemplos:

- Medicina: após uma bateria de exames preliminares, um sistema especialista pode indicar com considerável precisão se determinado tumor retirado de um paciente representa uma anomalia maligna ou benigna, antes mesmo do resultado da biópsia, fornecendo mais tempo de tratamento adequado aos pacientes. Pode ser utilizado também para indicar se um paciente é diabético, se um paciente possui problemas de coração, entre outros (Lin, 2007; Tsakonas, 2006).
- Sistema Bancário: pode ser utilizado para auxiliar na avaliação de clientes para liberação ou não de empréstimos, informando se o cliente possui características de bom pagador. Isso pode reduzir o número de empréstimos não quitados no banco (Eiben & Smith, 2003).

Os classificadores são definidos a partir de modelos, sendo um dos mais estudados, as Rede Neural Perceptron Multicamada - *Multilayer Perceptron* – MLP (Haykin, 1998). Concorrentemente, outras técnicas foram e ainda são desenvolvidas, com o intuito de fazer das máquinas verdadeiros bancos de conhecimento.

3
Revisão Bibliográfica

Este capítulo tem como objetivo apresentar alguns métodos conhecidos para a classificação oncológica de células cancerígenas e algumas técnicas já conhecidas na literatura, propostas para resolver o problema de Classificação de Padrões.

3.1 Classificação Oncológica de células cancerígenas

O auto-exame da boca é uma técnica simples para identificar possíveis anormalidades, como mudanças na aparência dos lábios e da parte interna da boca, endurecimentos, caroços, feridas, e inchaços. Entretanto, esse exame não substitui o exame clínico realizado por profissional experiente da área de saúde. E quando alguma lesão é identificada, o paciente ainda deve ser submetido a biópsias, que consistem em retirar amostras da lesão para determinar as características do tumor e o seu grau de malignidade (Ferreira, 2010).

Existem várias formas de obter esta amostra, como apresentado na Seção 2.2, e logo após coletado, este material é enviado ao laboratório de patologia e passa por um processo específico para a preparação de lâminas que serão analisadas por um patologista (Ferreira, 2010).

Para a classificação do grau de malignidade existem alguns métodos conhecidos. As classificações mais relevantes, validadas e citadas pela literatura estão apresentadas nas Seções 3.1.1 até 3.1.5 a seguir (Lourenço *et al.*, 2007).

3.1.1 Classificação Descritiva

Também conhecida como classificação histopatológica de Broders (1941), esta metodologia foi proposta inicialmente em 1920 e revisada em 1925, baseou-se no princípio fundamental de diferenciação celular descartando a história clínica

(Lourenço *et al.*, 2007). Foi dividida em quatro grupos diferenciados como mostrado na Figura 4.

Classificação Histopatológica de Broders	
Parâmetro	**Características**
Grau 1	0 a 25% de células indiferenciadas
Grau 2	25 a 50% de células indiferenciadas
Grau 3	50 a 75% de células indiferenciadas
Grau 4	75 a 100% de células indiferenciadas

Figura 4. Classificação de Broders (Lourenço *et al.*, 2007).

3.1.2 Sistema de Gradação Multifatorial

O principal objetivo foi propor uma técnica que não considerasse apenas a avaliação das células da lesão. Considera também a interface entre as células tumorais e o tecido hospedeiro (Lourenço *et al.*, 2007).

Foi proposta em 1984 por Anneroth e Hansen (1984), tendo como referência a classificação proposta por Jakobsson (1973) para carcinomas de laringe. Na Figura 5 pode-se observar o sistema de Gradação Multifatorial.

Sistema de Gradação Histopatológica Multifatorial (Anneroth)				
Relativo à população celular do tumor				
	PONTUAÇÃO			
Parâmetros Morfológicos	1	2	3	4
Estrutura	Lençóis sólidos e/ou configuração papilar	Fileiras e cordões	Pequenos grupos de células	Dissociação celular marcante
Padrão de ceratinização	Altamente ceratinizada	Ceratinização moderada	Ceratinização mínima	Sem ceratinização
Aberrações nucleares	Poucas	Moderadamente abundante	Poucos núcleos grandes e anaplásicos	Abundante e muitos núcleos grandes e anaplásicos
Número de mitoses*	Poucas (0-2)	Número moderado (3-4)	Numerosas (5-6)	Extremamente numerosas
Relativo à relação tumor-hospedeiro				
	PONTUAÇÃO			
Parâmetros Morfológicos	1	2	3	4
Modo de invasão	Membrana basal bem definida	Membrana basal menos nítida	Membrana basal indistinguível	Membrana basal indistinguível e invasão difusa
Estágio de invasão	Duvidosa ou microinvasão	Invasão nítida, mas envolvendo apenas a lâmina própria	Invasão abaixo da lâmina própria	Invasão massiva ampla e profunda
Resposta inflamatória	Marcante	Moderada	Leve	Nenhuma

Figura 5. Sistema de Gradação Histopatológica Multifuncional (Lourenço *et al.*, 2007).

3.1.3 Sistema da Gradação das Margens Invasivas

Proposta por Bryne (1989), baseia-se em modificações do sistema de graduação multifatorial, utilizando amostras provenientes de biópsias incisionais de tumores da mucosa bucal e alveolar.

Segundo Lourenço *et al.* (2007), os autores acreditavam que as células das áreas profundamente invasivas mostravam alterações mais parecidas com aquelas observadas em metástases e possuíam maior probabilidade de causar a disseminação do tumor, portanto, seriam consideradas como as células que iriam ditar o comportamento clínico da lesão maligna. Outra alteração foi a remoção do parâmetro estágio de invasão.

Essa classificação propunha pontuações para cada uma das características, sendo ao final, todos os cinco pontos somados, fornecendo o grau de malignidade

da lesão (Lourenço *et al.*, 2007). Na Figura 6 é apresentada a classificação histopatológica de Bryne (1989).

Sistema de Gradação das Margens Invasivas				
	PONTUAÇÃO			
Característica Morfológica	1	2	3	4
Grau de Ceratinização	Altamente ceratinizado (>50% das células)	Moderadamente ceratinizado (20-50% das células)	Ceratinização mínima (5-20% das células)	Sem ceratinização (0-5% das células)
Pleomorfismo Nuclear	Pouco pleomorfismo nuclear (>75% das células maduras)	Moderadamente abundante pleomorfismo nuclear (50-75% das células maduras)	Abundante pleomorfismo nuclear (25-50% das células maduras)	Extremo pleomorfismo nuclear (0-25% das células maduras)
Padrão de Invasão	Compressivo, bordas infiltrantes bem delineadas	Infiltrante, cordões sólidos, bandas ou fios	Pequenos grupos ou cordões de células infiltrantes (n>15)	Marcante e disseminada dissociação em grupos pequenos e/ou em células individuais
Infiltrado Linfo - plasmocitário	Marcante	Moderado	Leve	Ausente

Figura 6. Sistema de Gradação das Margens Invasivas (Lourenço *et al.*, 2007).

3.1.4 Avaliação Histopatológica de Risco

Está classificação foi proposta inicialmente em 2005, onde Brandwein (2005) afirmava que a condição das margens tumorais, antes considerada de grande influência, não possuía realmente valor preditivo (Lourenço *et al.*, 2007).

Nesta metodologia foram definidas três variáveis e para cada uma delas foram atribuídos valores numéricos. Ao final da análise, são somados os pontos atribuídos a todas as características, determinando assim, a classificação. O quadro da Figura 7 apresenta a Avaliação Histopatológica de Risco.

Avaliação Histopatológica de Risco para o CCE em boca (Brandwein-Gensler et al.)			
	Valores Atribuídos		
Variável Histopatológica	0	1	3
Invasão perineural	Nenhuma	Pequenos nervos	Grandes nervos
Infiltrado linfocítico	Contínuo	Grandes agregados	Pouco ou nenhum
Pior padrão de invasão	Padrão 1, 2 ou 3	4	5
Pontuação de Risco (Soma dos Pontos)	**Risco de recorrência local**	**Probabilidade de sobrevida total**	**Indicação para radioterapia adjuvante**
0	Baixo	Boa	Não
1 ou 2	Intermediário	Intermediária	Não
3 a 9	Alto	Pobre	Sempre

Figura 7.Avaliação Histopatológica de Risco (Lourenço *et al.*, 2007).

3.1.5 Gradação Histopatológica da OMS

Proposta em 2005 pela Organização Mundial da Saúde (OMS) baseou-se no grau de diferenciação celular e permitiu a classificação em três categorias (Lourenço *et al.*, 2007):

- Bem diferenciados – possuem a arquitetura tecidual semelhante a um padrão normal de epitélio escamoso.
- Moderadamente diferenciadas – apresentam certo grau de pleomorfismo nuclear, atividade mitótica e pouca ceratinização.
- Pouco diferenciados – Predomínio de células imaturas, numerosas mitoses típicas e atípicas, mínima ceratinização.

Gradação Histopatológica – OMS	
Parâmetros	**Características**
Pouco diferenciados	Predomínio de células imaturas Numerosas mitoses típicas e atípicas Mínima ceratinização
Moderadamente diferenciados	Certo grau de pleomorfismo nuclear e atividade mitótica Pouca ceratinização
Bem diferenciados	Arquitetura tecidual semelhante ao padrão normal do epitélio escamoso

Figura 8. Classificação recomendada pela OMS (Lourenço *et al.*, 2007).

3.2 Classificação de Células Cancerígenas utilizando Inteligência Artificial

Os sistemas classificadores são sistemas desenvolvidos pela área de IA que apresentam uma forma de preservar e disponibilizar o conhecimento de um determinado domínio proporcionando um diferencial competitivo a quem os possui. Esses sistemas buscam adquirir conhecimento possibilitando ao computador exercer funções antes desempenhadas pelos seres humanos com a utilização de conhecimento e raciocínio (Rezende *et al.*, 2003).

Rezende *et al.* (2003; pag.53, 54) apresentam várias técnicas desenvolvidas com o intuito de tornar mais efetivo os processos de aquisição do conhecimento, podendo estes, ser classificados em:

- Manuais – Esses sistemas utilizam um novo tipo de profissional especializado, denominado Engenheiro do Conhecimento. Este profissional, com a ajuda de especialistas da área a ser estudada, possui a tarefa de elaborar manualmente um modelo do domínio e das tarefas que requerem especialização. Esse processo é considerado um gargalo na construção de Sistemas Inteligentes.
- Semi-automáticos – Procuram minimizar a necessidade de um Engenheiro do Conhecimento, fornecer ferramentas aos especialistas que possibilitam a criação dos sistemas.
- Automáticos – Os sistemas de aquisição de conhecimento automáticos procuram obter, através de análise de dados já conhecidos, o conhecimento necessário.

Logo abaixo, na Seção 3.2.1 até 3.2.3, estão apresentados alguns métodos de aquisição do conhecimento automáticos presentes na literatura.

3.2.1 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) foram propostas inicialmente em 1943, onde Warren McCulloch e Walter Pitts apresentaram o primeiro modelo de neurônio

artificial (Bettiollo, 2009). Elas são utilizadas em varias aplicações e nas mais diferentes áreas: Matemática, Computação, Engenharia, Economia, Medicina, Psicologia, entre outras.

As RNAs correspondem a modelos matemáticos semelhantes às redes neurais biológicas e possuem a capacidade computacional adquirida por meio de algoritmos de aprendizagem e generalização. (Braga *et al.*, 2000; Haykin, 1994). São sistemas compostos por unidades com processamento simples, interligados entre si através de conexões ponderadas por pesos, que executam operações em paralelo e de forma distribuída (Haykin, 1999). Essas unidades com processamentos simples correspondem aos neurônios artificiais como mostrado na Figura 9.

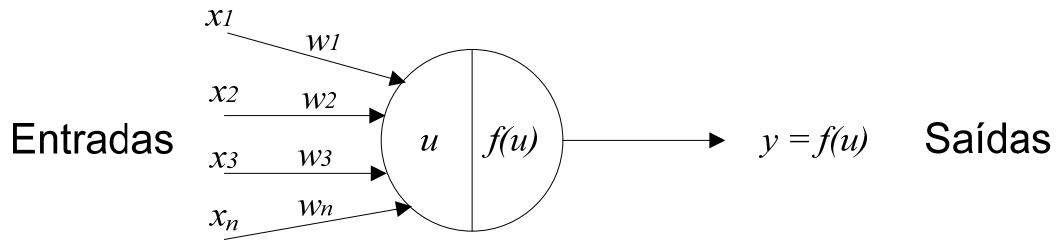

Figura 9. Visão esquemática de um neurônio artificial (Rezende *et al.*, 2003).

Rezende *et al.* (2003, pág. 142, 143) detalham os neurônios da seguinte forma: as entradas correspondem ao vetor $X = [x_1, x_2, x_3, .., x_n]^T$ e possuem a dimensão n. Para cada uma das entradas x_i corresponde um peso w_i, e a soma das entradas x_i ponderadas pelos pesos correspondentes w_i, chamada de saída u, corresponde a $u = \sum_i w_i x_i$. Já a saída y, chamada de saída de ativação, é obtida pela aplicação de uma função $f(.)$ à saída linear u. Essa função $f(.)$ pode assumir varias formas, geralmente não-lineares, como na Equação 1, onde o θ é o limiar ao qual a saída é ativada. Algumas das funções de ativação mais utilizadas são: função sinal, função rampa, a sigmoidal e a tangente hiperbólica (Prechelt, 1994; Haykin, 1994).

$$f(u) = \begin{cases} 0 & u < \theta \\ 1 & u \geq \theta \end{cases} \quad (1)$$

Como podemos observar cada neurônio executa uma função simples, e a RNA como um todo tem capacidade para problemas mais complexos (Braga *et al.*, 2000). Logo a seguir, na Figura 10, é apresentado um modelo de rede neural com 4 entradas, 2 saídas e 3 neurônios na camada intermediária.

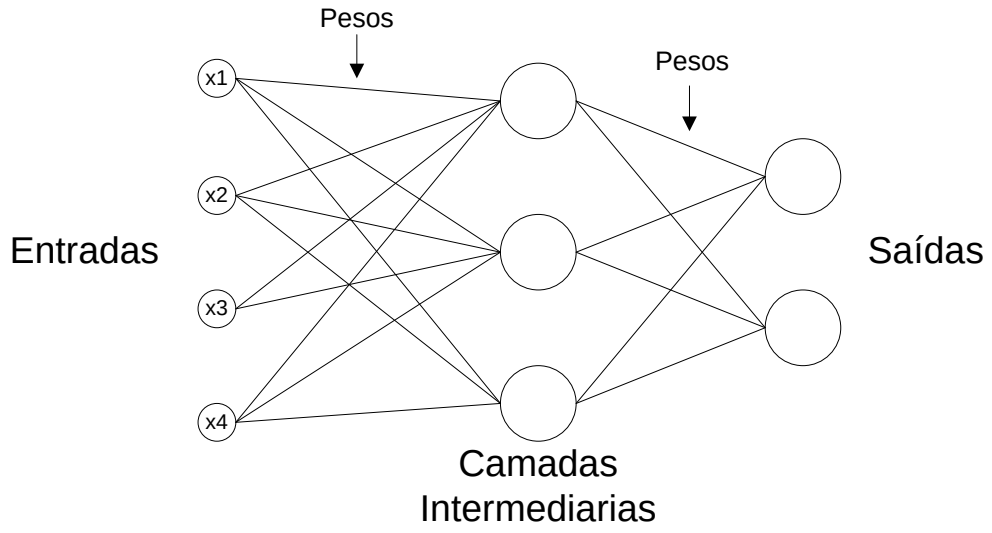

Figura 10. RNA com 4 entradas, 2 saídas e 3 neurônios na camada intermediária.

Segundo Silva (2005), "*A maneira pela qual os neurônios de uma rede estão estruturados está intimamente ligada com o algoritmo de aprendizagem usado para treinar a rede*" e os tipos de treinamentos existentes podem ser consultados em Haykin (2001).

Basicamente, são identificadas três formas de arquiteturas de rede fundamentalmente diferentes (Silva, 2005):

- Redes Alimentadas com Camada Única – é composta pela camada de entrada de dados e seguida de apenas uma única camada de processamento que calcula e fornece as saídas.

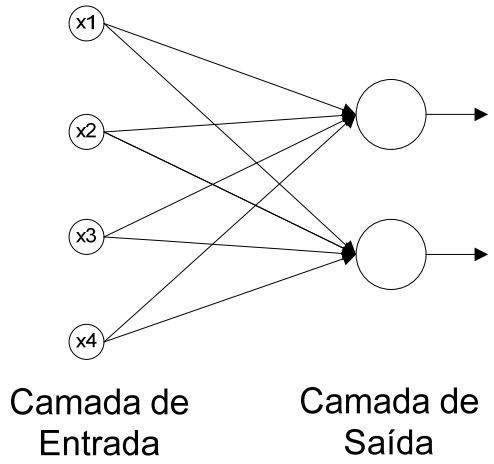

Figura 11. RNA alimentada com camada única

- Redes Alimentadas com Múltiplas Camadas – também conhecida por *Perceptrons* de Múltiplas Camadas apresentam uma ou mais camadas ocultas.

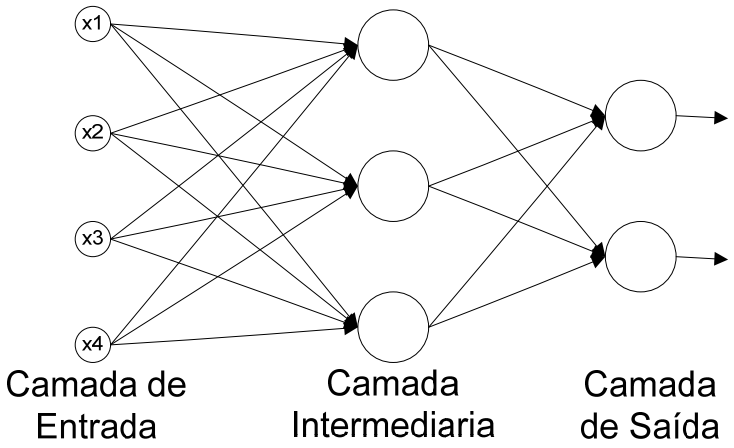

Figura 12. RNA com Mútiplas Camadas.

- Redes Recorrentes – distingue da rede com múltiplas camadas por ter pelo menos um laço de realimentação.

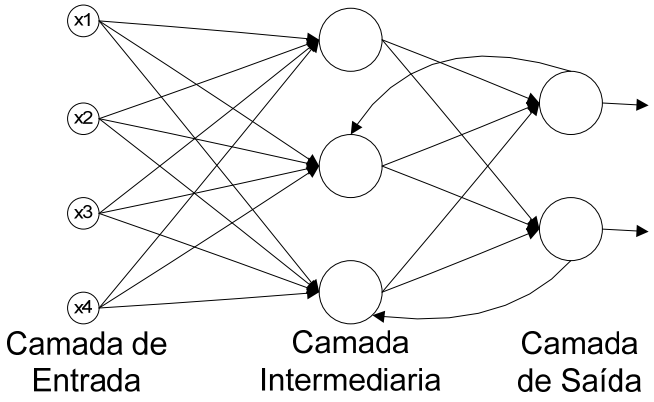

Figura 13. RNA recorrente.

Após determinarmos a organização da rede, as configurações de quantidade de entradas, camadas intermediárias e quantidade de saídas, cabe a RNA encontrar a melhor solução de configuração para os pesos e para o limiar da função de ativação. Esta etapa é denominada aprendizagem, e é feita baseado em exemplos reais apresentados a rede.

Encerrada a etapa de aprendizagem, a rede estará pronta para realizar novas classificações.

3.2.2 Algoritmos Genéticos

Os Algoritmos Genéticos (AGs) enquadram-se em um ramo da IA denominado Computação Evolutiva e foi proposto inicialmente por John Holland (1975). São algoritmos evolucionários baseados nos mecanismo naturais de sobrevivência e reprodução das populações, onde os indivíduos mais adaptados ao seu ambiente sobrevivem e reproduzem a taxas mais altas do que indivíduos menos adaptados (Souza, 2004; Almeida, 2007).

Almeida (2007) e Rezende *et al.* (2003) afirmam que ao empregarmos os AG em problemas do mundo real, cada indivíduo da população, denominado cromossomo, corresponde a uma possível solução para o problema. Esses indivíduos são modelados como estruturas de dados fixas, normalmente definidas por sequências de *bits*, denominados genes, que representa a presença (1) ou ausência (0) de uma determinada característica. Desta forma os elementos podem ser combinados formando as características reais do indivíduo. Na Figura 14 podemos observar um exemplo de cromossomo e gene.

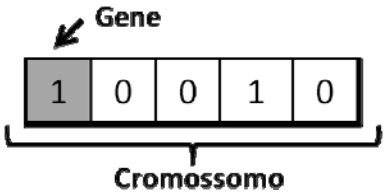

Figura 14. Exemplo de Cromossomo do Algoritmo Genético

Os AG possuem um conjunto de passos distintos e bem definidos, sendo que cada passo possui muitas variações possíveis. O funcionamento básico do AG descrito em Rezende *et al.* (2003, pág. 229) é:

> *"Inicialmente, é gerada uma população formada por um conjunto aleatório de indivíduos que podem ser vistos como possíveis soluções do problema.*
>
> *Durante o processo evolutivo, a população é avaliada: para cada indivíduo é dada uma nota, ou índice, refletindo sua habilidade de adaptação a determinado ambiente. Uma porcentagem dos mais adaptados é mantida, enquanto os outros são descartados (darwinismo). Os membros mantidos pela seleção*

> *podem sofrer modificações em suas características fundamentais por meio de mutação e cruzamento (crossover) ou recombinação genética gerando descendentes para a próxima geração. Esse processo, chamado de reprodução, é repetido até que uma solução satisfatória seja encontrada."*

Encontrando a melhor configuração de genes do cromossomo, o algoritmo está pronto para realizar novas classificações.

3.2.3 Programação Genética

O paradigma da Programação Genética (PG) pertence à família dos Algoritmos Evolucionários e consiste de uma técnica que propicia a geração automática de programas de computadores, ou seja, o objetivo é encontrar um programa de alta qualidade no espaço de todos os programas possíveis (Almeida, 2007). Foi proposta, inicialmente, por Koza (1992) e inspira-se na idéia da teoria da evolução de Darwin (1859) e na genética.

Na PG, diferente dos AG, os indivíduos de uma população são programas de computadores geralmente representados por Árvores Sintáticas, como mostradas na Figura 15, e a avaliação dos indivíduos ocorre através da execução do programa representado por este indivíduo (Rezende *et al.*, 2003).

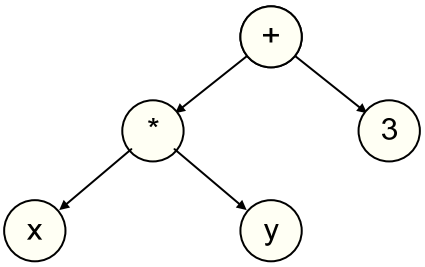

Figura 15. Árvore representando o programa para calcular (x*y)+3.

O algoritmo básico da PG consiste em (Koza, 1992):

- O primeiro passo do algoritmo é a criação de uma população de indivíduos aleatoriamente;
- Após a criação da população inicial são executados, até atingir o critério de parada, os seguintes passos:

- Realiza uma avaliação da adaptação de cada indivíduo;
- Baseado na avaliação é aplicado um método de seleção para escolher alguns indivíduos;
- Modifica-se os indivíduos escolhidos através de operadores genéticos, como mutação ou cruzamento, criando assim uma nova população.

O critério de parada é atingido quando o sistema já encontrou uma forma de realizar a classificação. Mais detalhes sobre as técnicas da PG podem ser analisados na Seção 4.2.

4
Método Proposto

Esta seção tem por objetivo descrever a nova técnica de classificação proposta pelo presente trabalho. A Seção 4.1 apresenta detalhes de sistemas classificadores, na Seção 4.2, estão descritos fundamentos teóricos da PG, e na Seção 4.3 uma descrição do sistema desenvolvido nesta pesquisa.

4.1 Classificação de Padrões

Rezende *et al.* (2003) descrevem a aprendizagem de máquina como o desenvolvimento de técnicas que possibilitam ao computador adquirir conhecimento de forma automática baseando em aprendizagem indutiva. Essa aprendizagem indutiva utiliza um conjunto de exemplos, para os quais o rótulo da classe associada é conhecido, com a finalidade de obter conclusões genéricas sobre esse conjunto particular de exemplos. Geralmente, uma boa indicação para o uso destas técnicas é a existência de um profissional capaz de solucionar o problema.

Existem duas formas de obtenção da aprendizagem indutiva:

- Supervisionado: neste método é fornecido um conjunto de exemplos e para cada exemplo é descrito um vetor de valores de características e o rótulo da classe correspondente. O objetivo do algoritmo de aprendizagem é construir um classificador que possa determinar a classe correspondente a novos exemplos fornecidos ao sistema;
- Não-supervisionado: neste método os exemplos fornecidos não indicam o rótulo da classe correspondente, desta forma, o algoritmo de aprendizagem tenta determinar se alguns deles podem ser agrupados de alguma forma.

Em casos de sistemas supervisionados, se os rótulos das classes forem discretos, o problema é conhecido como *classificação*, e se forem valores contínuos como *regressão* (Rezende *et al.*, 2003, pág. 91).

O funcionamento dos sistemas classificadores geralmente é dividido em etapas, sendo estas:

- Primeira Etapa (Elaboração da base de dados): Nesta fase, Figura 16, um especialista analisa os padrões com o intuito de determinar as características mais importantes para distinguir uma classe de outra (nesta pesquisa os padrões são as amostras de células de câncer de boca). Este processo é feito por um ser humano, especialista no problema, e é responsável pela elaboração de uma base de dados previamente conhecida que será transmitida ao sistema na fase de aprendizagem. Essa base contém informações de vários padrões, e para cada padrão é informado um vetor de características e a sua respectiva classe. A elaboração desta base deve ser feita com cautela, pois se o número de exemplos for insuficiente, ou se os exemplos não forem bem escolhidos, o método de classificação encontrado pode ser de pouco valor;

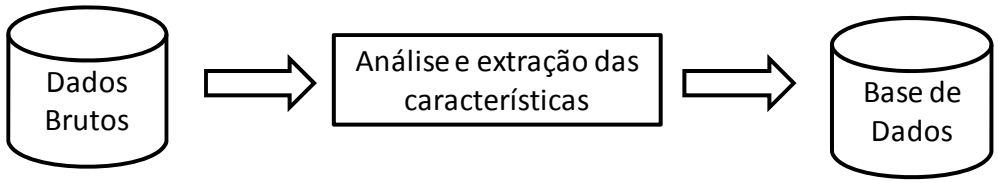

Figura 16. Etapa de elaboração da base de dados.

- Segunda Etapa (Aprendizagem): Este processo pode ser analisado na Figura 17, e consiste basicamente na utilização de técnicas de IA que possibilitem ao sistema adquirir conhecimento, por meio de exemplos já conhecidos do problema a ser resolvido. Algumas das técnicas utilizadas são RNA, PG ou AG descritas na Seção 3.2. Os exemplos já conhecidos são obtidos através das bases de dados elaboradas na primeira etapa e o conhecimento adquirido pelo sistema deve ser armazenado em uma base de conhecimento, podendo ser utilizado em futuras classificações. As técnicas de aprendizagem geralmente começam sem nenhum conhecimento, e através de algoritmos de treinamento buscam soluções mais adaptadas para realizar a classificação. Em geral, as regras de classificação são expressas na forma "SE a ENTAO b". O "a" pode

ser formado por um conjunto de condições ou uma expressão e o "b" representa o resultado da classificação;

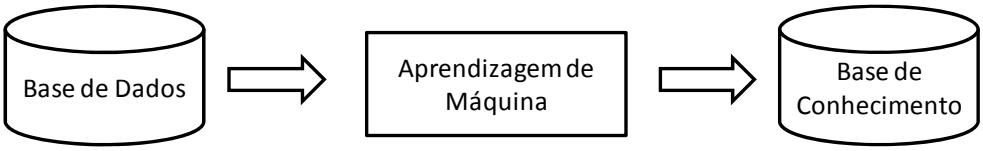

Figura 17. Etapa de aprendizagem de máquina

- Terceira Etapa (Classificação de novos padrões): Esta fase consiste em classificar padrões aos quais ainda não é conhecida a classe correspondente. O especialista deve extrair as características do padrão que são utilizadas pelo sistema, essas características são processadas utilizando a base de conhecimento adquirida na etapa anterior e o sistema de classificação retorna a classe correspondente ao padrão (Figura 18).

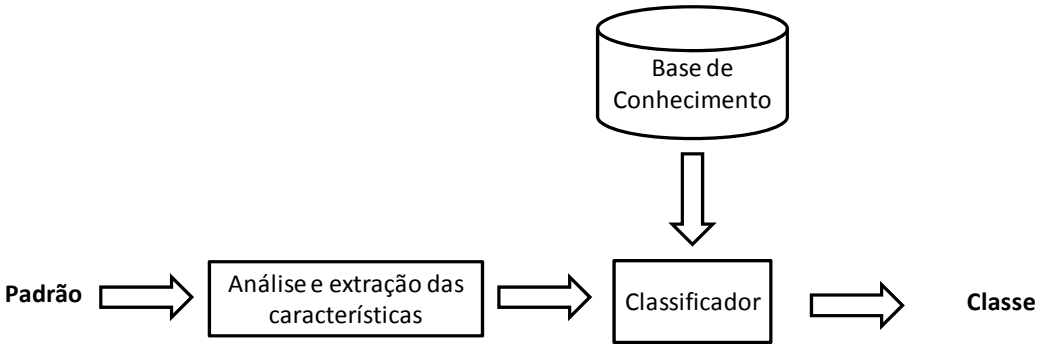

Figura 18. Classificação de novos padrões.

O funcionamento básico de sistemas classificadores foi detalhado nesta seção, sendo a seção seguinte, utilizada para apresentar uma técnica de aprendizagem de máquina, já conhecida na literatura, utilizada para a construção automática de programas de computadores.

4.2 Programação Genética

A PG consiste em uma abordagem para geração automática de programas de computadores a partir de uma descrição em alto nível do problema a ser atacado

(Rodrigues, 2007; Rezende *et al.*, 2003). Foi desenvolvida por John Koza (1992) e é baseada na idéia de seleção natural de Darwin (1859), onde indivíduos mais aptos terão mais descendentes do que os menos aptos ao longo de todo processo evolutivo (Almeida, 2007).

Os algoritmos de PG são baseados em populações, sendo essas populações, um conjunto de indivíduos. Cada um desses indivíduos representa um programa de computador ou uma possível solução do problema e possui um conjunto de características próprias que os distingue dos demais indivíduos da população (Almeida, 2007).

A representação dos indivíduos em PG utiliza normalmente uma estrutura em árvore denominada Árvores Sintáticas, proposta inicialmente por Koza (1992). Essas árvores podem representar sem dificuldades sistemas formais recursivos, como fórmulas aritméticas. Podemos exemplificar esta representação com a função $g(x,y) = x*y+3$ na Figura 19.

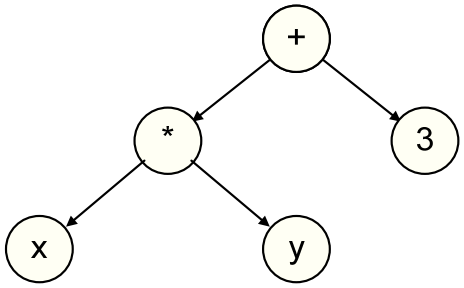

Figura 19. Representacão da função g(x,y) = x*y+3.

Para evitar que indivíduos incorretos (programas que não possam ser compilados) sejam criados, recentemente foi introduzido uma abordagem de PG orientada a gramática (Whigham, 1996; Wong, 2000). Essas gramáticas auxiliam o processo de criação aleatória de indivíduos, facilitando a geração de programas válidos (Minku *et al.,* 2003). São compostas por quatro componentes:

- *V* – Conjunto de Símbolos não terminais
- *T* - Conjunto de Símbolos terminais
- *R* – Conjunto de Regras
- *P* – é um símbolo não terminal conhecido como símbolo de partida ou símbolo inicial.

É formalmente descrita na seguinte fórmula:

$$GRAMATICA = \{V, T, R, P\} \quad (1)$$

Na Seção 4.2.1 é apresentado o meta-algoritmo básico da PG, sendo as seções seguintes, utilizadas para detalhar melhor os passos do meta-algoritmo.

4.2.1 Algoritmo da Programação Genética

O meta-algoritmo da PG, detalhado pelo fluxograma da Figura 20, começa seu processamento com a criação de uma população inicial formada por um conjunto aleatório de indivíduos (possíveis soluções do problema). Após ser criada, a condição de parada é verificada, se atendida, o melhor individuo é apresentado e o algoritmo é encerrado, caso contrário, ocorre uma serie de repetições onde são criadas novas populações até que seja satisfeita essa condição de parada.

A criação dos indivíduos para a nova população consiste em:

- Passo1 – seleciona o operador a ser utilizado, este operador nada mais é do que a forma pela qual o novo indivíduo será criado, mutação ou cruzamento. Se o operador escolhido for *op1*, ou seja, operador de mutação, o algoritmo realiza o passo2, caso contrário, ocorre o cruzamento e o algoritmo realiza o passo3.
- Passo 2 - seleciona um indivíduo da população existente, realiza a mutação deste indivíduo e o adiciona na nova população.
- Passo3 - seleciona dois indivíduos da população existente, realiza o cruzamento, e adiciona os indivíduos criados na nova população.

As populações em PG possuem tamanho fixo, e após criada uma nova população, a antiga é descartada.

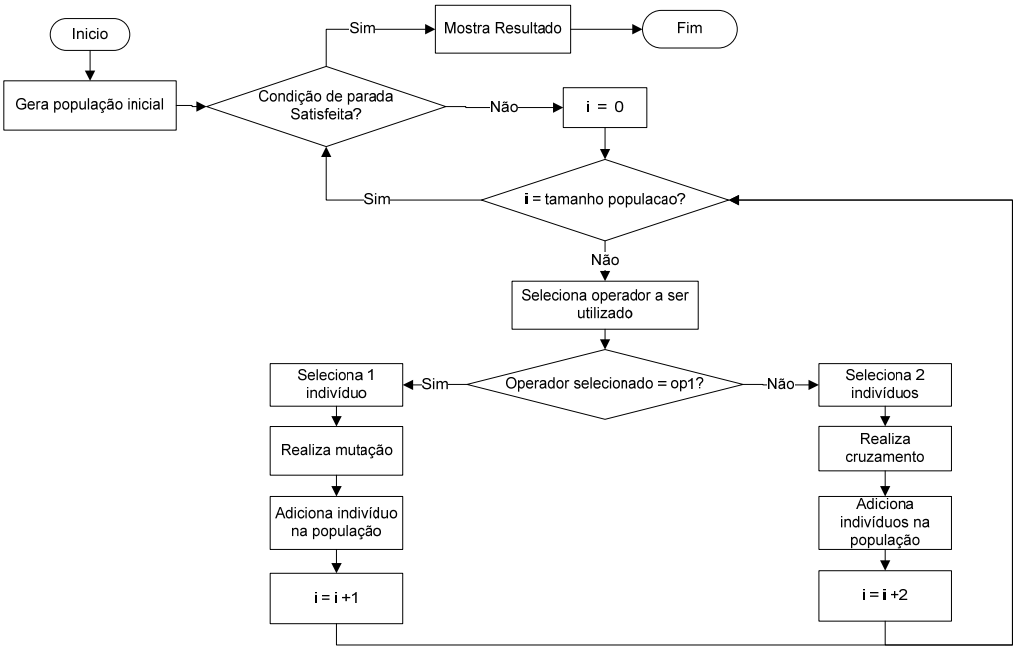

Figura 20. Estrutura básica do algoritmo de Programação Genética (Koza, 1992, p. 76).

4.2.2 Iniciando o algoritmo (Criação da primeira população)

O primeiro passo do algoritmo básico da PG é criar a população inicial. Nesta etapa, geralmente, nenhuma heurística é usada, sendo a criação conduzida de forma aleatória. Alguns métodos utilizados para criação das árvores podem ser encontrados em Koza (1998) e Almeida (2007):

- *Full method*: Neste algoritmo é definido um parâmetro D_{max} que representa a altura da árvore a ser criada, desta forma, todos os nós-folha da árvore possuem a mesma distância até o nó-raiz criando sempre árvores completas;
- *Grow method*: Neste algoritmo são criadas árvores de profundidade D_{max}, podendo existir nós-folhas cuja distância até o nó-raiz seja menor que D_{max};
- *Ramped-half-and-half* – consiste em uma combinação balanceada do *Full method* e do *Grow method*, sendo cada um dos métodos escolhidos 50% das vezes.

4.2.3 Avaliação da Qualidade da População

Para avaliar o quão bom é um indivíduo, é necessário associar a cada um o valor de adaptabilidade correspondente (*fitness*), ou seja, um valor que represente a capacidade deste indivíduo em resolver um determinado problema. A forma de determinar este valor alterna de acordo com o tipo de retorno do sistema.

Uma das métricas utilizada para avaliar árvores sintáticas que representam expressões matemáticas é a Soma de quadrados dos erros (SQE), análise amplamente utilizada na arquitetura de diferentes Redes Neurais Artificiais. Mais detalhes podem ser obtidos em Haykin (1998). A métrica SQE é descrita na equação 2.

$$SQE = \sum_{i=1}^{n} \left(\hat{X}_i - X_i \right)^2 \qquad (2)$$

onde:

$\hat{X}_i$ – resultado esperado, ou seja, resultado que deve ser retornado da expressão, representa a classe ao qual o objeto pertence.

X_i – representa o número que realmente foi retornado da expressão.

Quanto menor o valor do SQE, mais adaptada é a solução encontrada.

4.2.4 Seleção

A seleção é necessária para escolher os indivíduos sobre os quais serão aplicados os operadores genéticos. É realizada utilizando os valores de adaptabilidade e consiste em escolher aqueles que irão sofrer transformações e consequentemente compor a nova geração (Almeida, 2007). Os principais métodos de seleção são:

- Seleção proporcional: Neste método, quanto maior for o *fitness* de um indivíduo, maior a chance de ser selecionado;
- Seleção por torneio: primeiro seleciona-se *t* indivíduos, sendo *t* o tamanho do torneio. Logo após, é escolhido o melhor entre os *t* indivíduos;

- Seleção Aleatória: a escolha dos pais ocorre de forma aleatória, sem a utilização de avaliação dos indivíduos.

4.2.5 Operadores Genéticos

A evolução da população em PG é obtida através da localização de indivíduos mais adaptados ao problema. A localização desses indivíduos ocorre através de operadores genéticos que causam modificações em indivíduos já conhecidos pela população. Os operadores mais conhecidos são:

- Mutação: A mutação na PG é a criação de um novo indivíduo através de pequenas alterações aleatórias em indivíduos já existentes na população. Existem varias técnicas de operadores de mutação disponíveis na literatura. A mutação de sub-árvores, ilustrada na Figura 21, é uma das mais conhecidas e ocorre da seguinte forma: tendo feita a seleção de um individuo existente, seleciona-se, aleatoriamente um nó deste indivíduo que passa ser a raiz de uma sub-árvore que deve ser substituída por uma nova árvore criada aleatoriamente (Eiben & Smith, 2003);

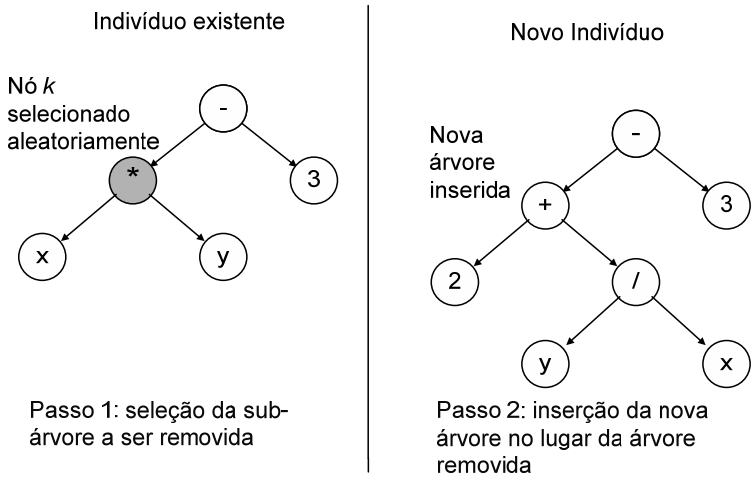

Figura 21. Operador genético de Mutação.

- Cruzamento: Já no cruzamento selecionam-se dois indivíduos pais, logo após, é selecionado um nó de cada pai e trocam-se entre eles as sub-árvores, cuja raiz são os nós selecionados, formando assim dois

filhos (indivíduos), com as características dos pais (Eiben & Smith, 2003), como mostrado na Figura 22.

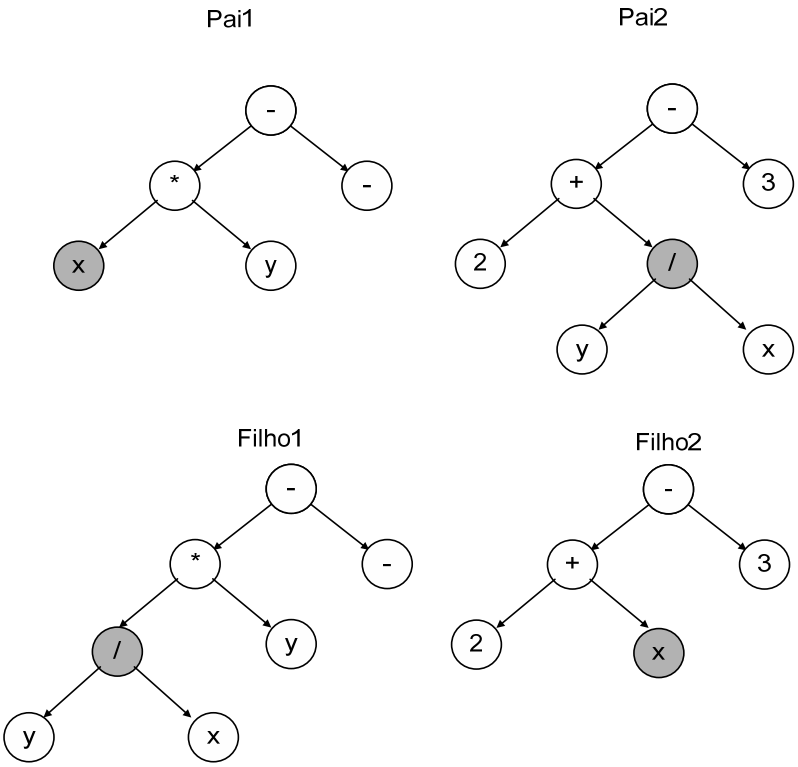

Figura 22. Operador genético de Cruzamento

4.2.6 Condição de Parada do algoritmo

A condição de parada é responsável pela finalização do algoritmo. Segundo Koza (1992), algumas das condições mais usadas são: atingir o número máximo de gerações ou atingir algum valor de adaptabilidade pretendido.

4.3 Sigm-Tree

A Técnica de Classificação proposta por este trabalho consiste em uma adaptação da Programação Genética. Basicamente, este classificador busca expressões aritméticas que sejam capazes de realizar a categorização de determinados objetos

em diferentes classes. Na classificação de células cancerígenas apresentada pela pesquisa, estas classes são definidas em: Grau I, Grau II e Grau III.

A técnica funciona da seguinte forma: para cada problema de classificação é localizado um conjunto de expressões, sendo cada expressão correspondente a uma classe do problema. Após ter definido este conjunto, são informados dados numéricos do objeto a ser analisado, sendo estes dados, processados por todas as expressões. A expressão que retornar o valor mais alto representa a classe escolhida.

Para representar as expressões são utilizadas Árvores Sintáticas como mostrado na Figura 23.

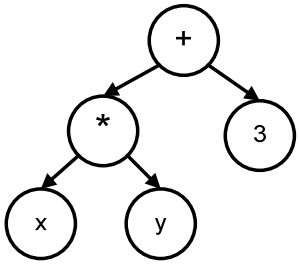

Figura 23. Representação da função g(x,y) = x*y+3.

Cada árvore armazena uma expressão que contém funções matemáticas (soma, subtração e multiplicação), constantes e características do objeto a ser identificado (mapeados em números reais). Essas características do objeto são fundamentais para a classificação, pois representam as características do objeto que queremos analisar.

O universo de expressões possíveis é definido através de uma Gramática Livre de Contexto aqui denominada de GAS. Na Equação 3 até a 6, utilizadas para detalhar a gramática, o *V* representa o conjunto de símbolos não terminais, *T* o conjunto de símbolos terminais, *R* o conjunto de regras e *P* o símbolo inicial (símbolo de partida).

$$GAS = \{V,T,R,P\}$$

$$V = < ROOT >,< SIGM >,< EXP >,< BINARY >,< UNARY >,< LEAF >,$$
$$< SUM >,< SUB >,< PROD >$$

$$T = \Re \cup \{s_1,s_2,s_3,...,s_n\} \cup \{(,)\}$$

$P = < ROOT >$

O conjunto de regras R é formado pelas fórmulas a seguir:

$< ROOT > \rightarrow < SIGM >$

$< SIGM > \rightarrow \frac{1}{(1 + e^{(-1 \times < EXP >)})}$

$< EXP > \rightarrow (< BINARY >) | (< UNARY >) | (< LEAF >)$

$< BINARY > \rightarrow < SUM > | < SUB > | < PROD >$

$< UNARY > \rightarrow < SIGM >$

$< LEAF > \rightarrow < CONST > | < VAR >$

$< SUM > \rightarrow < EXP > + < EXP >$

$< SUB > \rightarrow < EXP > - < EXP >$

$< PROD > \rightarrow < EXP > \times < EXP >$

$< CONST > \rightarrow \{y \mid y \in \Re\}$

$< VAR > \rightarrow s_1 \mid s_2 \mid ... \mid s_n$

Foi definido que toda árvore tem como raiz apenas a função sigmóide logística <SIGM>. Esta função garante que todos os valores retornados pelas expressões encontrem-se no intervalo entre 0 e 1. Desta forma, a expressão que retornar o valor mais próximo de 1 é a expressão correspondente a classe escolhida. Vale ressaltar aqui, que em casos de empate o sistema retorna o maior nível de malignidade, indicando um tratamento mais agressivo. A representação para esta função utilizada neste trabalho é definida por:

$$SIGM = \frac{1}{(1 + e^{(-1 \times < EXP >)})}$$

A sigmoide por sua vez pode abrigar qualquer expressão matemática dentro desta linguagem descrita pela gramática, incluindo os operadores binários (multiplicação <PROD>, soma <SUM> e subtração <SUB>) e constantes. As constantes estão sempre presentes nas folhas das Árvores Sintáticas e são divididas

em duas classes: constantes reais <CONST>, e constantes que representam características do padrão analisado <VAR> (vetor s).

O não terminal <VAR> é fundamental para a classificação dos padrões. No exemplo apresentado na Seção 2.3, que identifica se uma pessoa é do sexo masculino ou do sexo feminino, esta regra implicaria nos seguintes componentes: *s1* sendo o tamanho, *s2* sendo peso, *s3* sendo comprimento do cabelo e *s4* o tom de voz. Portanto, as regras do não terminal <VAR> representam as características dos objetos e a sua cardinalidade (*n*) corresponde à quantidade de características conhecidas.

Nos primeiros estágios desta pesquisa, foram analisadas adaptações desta linguagem com o intuito de aumentar o poder de classificação do método estudado, sendo observado que uma gramática simples possui o mesmo potencial das demais, porém não é necessário tratar exceções presentes na divisão por 0, por exemplo.

4.3.1 Estrutura do Algoritmo

Antes de detalharmos o algoritmo proposto, é necessário saber que para executar o classificador desenvolvido, é preciso uma base de dados com dados de vários objetos que desejamos aprender a classificar, e para cada objeto deve ser informado também a sua classe correspondente. Esses objetos devem ser divididos em diferentes conjuntos:

- Conjunto de treinamento: Utilizado no processo de aprendizagem do algoritmo. Auxilia na busca por árvores bem adaptadas, sendo usado para avaliar as expressões encontradas;
- Conjunto de validação: Utilizado para determinar o momento de parada do algoritmo de busca de novas árvores, além de ser usado para escolher a árvore a ser retornada no final do algoritmo;
- Conjunto de teste: Utilizado para avaliar a generalização das fórmulas encontradas para classificação.

O fluxograma básico do algoritmo proposto pode ser observado na Figura 24.

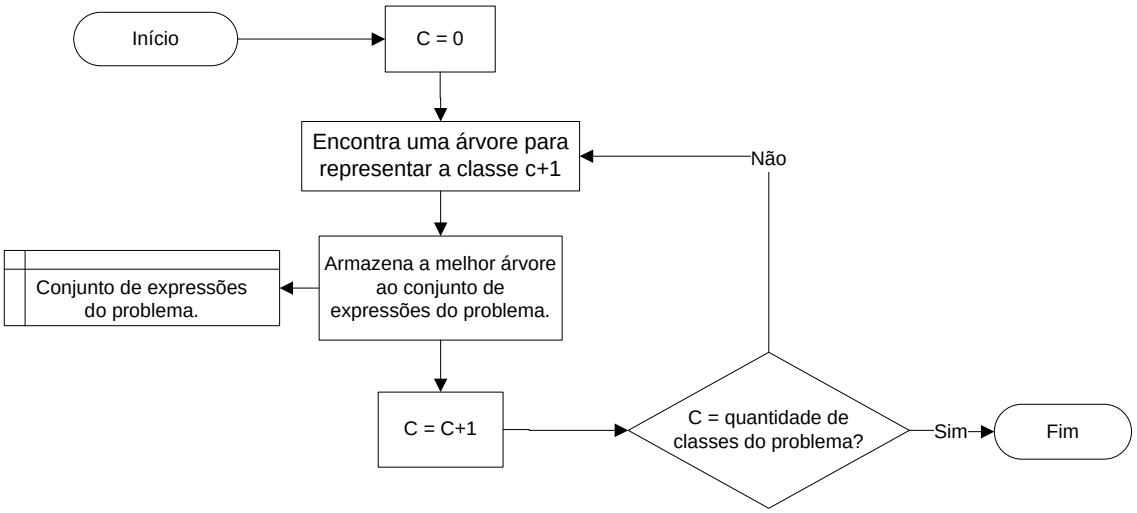

Figura 24. Fluxograma do algoritmo proposto.

O primeiro passo do algoritmo consiste em atribuir o valor 0 a variável C. Esta variável garante a criação de uma expressão para cada classe do problema e o valor 0 indica que ainda não foi criada nenhuma delas. Logo após, é encontrada uma expressão para representar a primeira classe do problema (classe 1), sendo esta, armazenada no conjunto de árvores do classificador, e a variável C é acrescida de 1. Ao final destes passos, é verificado se o número de árvores encontradas é igual ao número de classes do problema, em caso positivo, o algoritmo é finalizado, caso contrário, o algoritmo entra em um laço de repetições até que esta condição seja atingida. Esse laço de repetições consiste em:

- Criar uma árvore correspondente a próxima classe;
- Incluir essa nova árvore ao conjunto de árvores do classificador;
- Incrementar 1 a variável C.

A técnica utilizada para localizar a expressão correspondente a cada classe é descrita no fluxograma representado pela Figura 25.

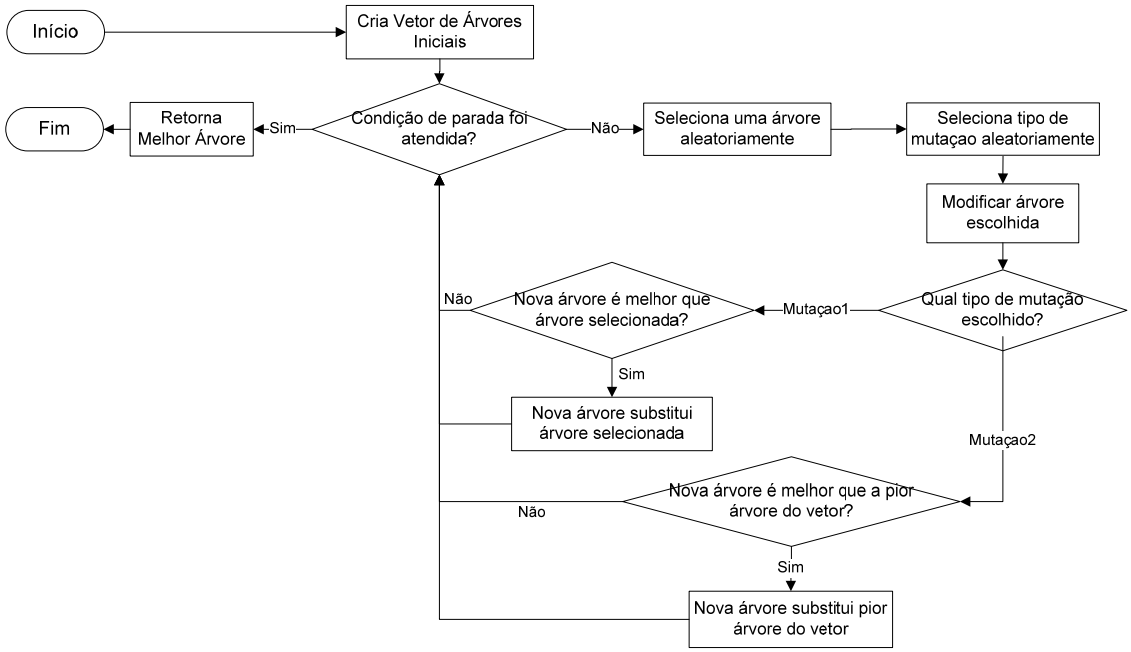

Figura 25. Algoritmo utilizado para encontrar novas árvores.

Inicialmente é criado um conjunto de árvores denominado população. No passo 2, a condição de parada é verificada, se atendida, a melhor árvore é apresentada e o algoritmo é encerrado, caso contrário, ocorre um laço de repetições:

- Passo 3: seleciona uma árvore aleatoriamente, esta árvore selecionada é denominada árvore pai;
- Passo 4: seleciona um operador de mutação para modificar a árvore pai;
- Passo 5: cria uma cópia da árvore pai e a modifica, criando assim uma nova árvore;
- Passo 6: Verifica qual tipo de mutação foi utilizada. Se a mutação utilizada for Mutação 1 o algoritmo pula para o passo 7, caso contrário, pula para o passo 8;
- Passo 7; Se a nova árvore for melhor do que a árvore pai, a árvore pai é substituída pela nova árvore que passa a fazer parte da população, caso contrário, a nova árvore é descartada e a árvore pai continua fazendo parte da população. Logo após o algoritmo pula para o passo 9;

- Passo 8: Encontra-se a pior árvore da população, se ela for pior que a nova árvore, ela é descarta e a nova árvore passa a fazer parte da população, caso contrário, a nova árvore é descartada. Logo após o algoritmo pula para o passo 9;
- Passo 9: Depois de avaliar a nova árvore e colocá-la ou não na população, é verificado novamente se a condição de parada foi atingida.

Este laço de repetição ocorre ate que essa condição de parada seja atingida. Abaixo todas as etapas do fluxograma são descritas com mais detalhes.

4.3.2 Criação do vetor de Árvores Inicial

O primeiro passo do algoritmo é a criação das expressões iniciais, ou seja, a criação de um vetor de árvores que representem a população inicial do sistema. O método utilizado para criação dessas árvores é o *Full method*, já detalhado na Seção 4.2.2, e o valor de D_{max} (altura das árvores) foi definido como 3.

4.3.3 Avaliação da Qualidade das Árvores Sintáticas (Função Objetivo)

Para avaliação da árvore sintática, são consideradas duas informações:

- Soma dos Erros Quadráticos (SQE): Função apresentada na seção 4.2.3, quanto menor o valor do SQE, melhor é a expressão encontrada;
- Quantidade de acerto: Número de objetos que a fórmula foi capaz de classificar corretamente, quanto maior a quantidade de acerto melhor a árvore.

Utilizando esses dados, foram definidas três formas diferentes de analisar a melhor árvore:

- Apenas pelo valor do SQE;
- Apenas pela quantidade de acerto;

- Analisando a quantidade de acerto e valor do SQE juntos.

Vale ressaltar que o método a ser utilizado deve ser definido pelo usuário do sistema durante a sua execução.

4.3.4 Encontrando novas Árvores

Assim como todos os outros sistemas de classificação, o classificador descrito neste trabalho está sujeito ao erro, que deve ser minimizado. Neste trabalho, a minimização será focada na localização de fórmulas mais adequadas para a classificação.

Para localizar novas fórmulas, uma árvore da população é escolhida aleatoriamente. A criação da nova árvore pode ser feita de duas formas:

- Mutação 1 (Mutação Simples): O algoritmo de vizinhança cria um clone da árvore selecionada, escolhe um nó *k* aleatoriamente desta árvore clone e substitui este nó por um nó do mesmo tipo, ou seja, se o nó escolhido for uma função matemática, é selecionada, aleatoriamente, outra função matemática para substituí-la, se o nó for uma variável ou constante, é selecionada uma variável ou uma constante, aleatoriamente, para substituí-la. Um exemplo deste operador pode ser visto na Figura 26;

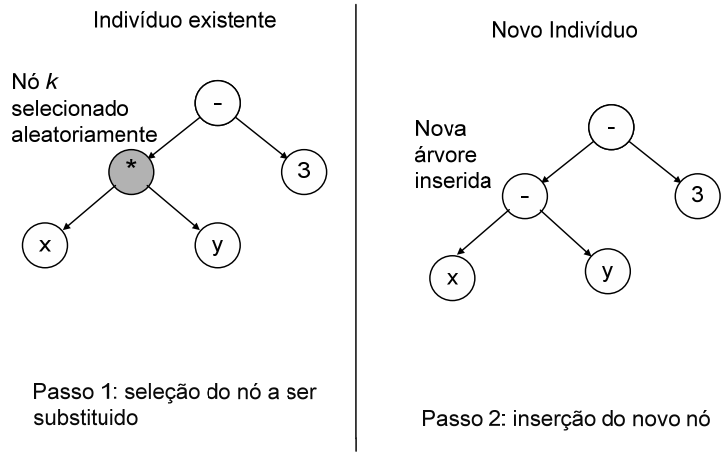

Figura 26. Operador genético de Mutação1

- Mutação 2: O algoritmo cria um clone da árvore selecionada, escolhe um nó *k* da árvore clone aleatoriamente e substitui a sub-árvore com

raiz *k* por uma nova árvore criada pelo Algoritmo *Grow method* ou pelo Algoritmo *Full method*. Para o parâmetro D_{max} foi definido o valor 3 e mais detalhes sobre estes métodos podem ser encontrados na seção 4.2.2.

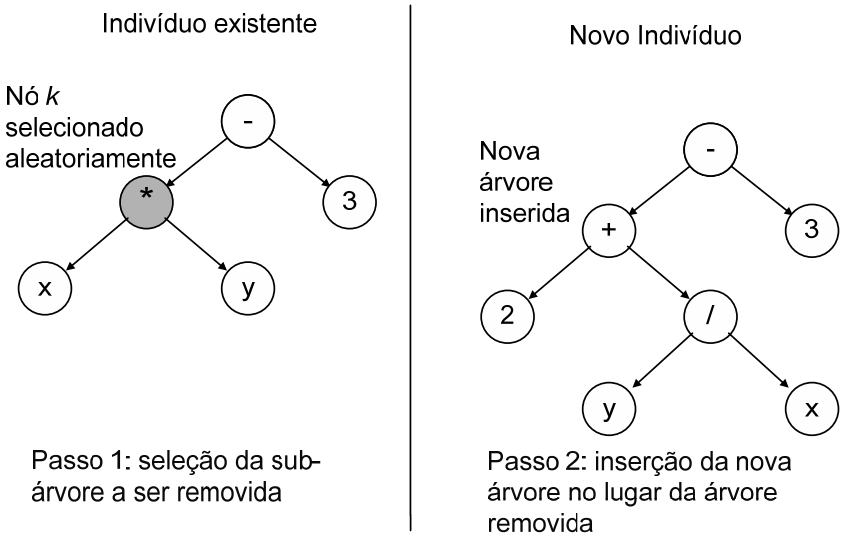

Figura 27. Operador genético de Mutação2.

Esta técnica não utiliza operadores de cruzamento e a escolha do método de mutação a ser utilizado é efetuada aleatoriamente pelo algoritmo.

Após a criação da nova árvore, esta passa por um processo de avaliação utilizando o conjunto de treinamento da base de dados. Esta avaliação varia de acordo com o processo escolhido pelo usuário do sistema, mais detalhes na Seção 4.3.3.

4.3.5 Condição de Parada do algoritmo de busca

A parada do algoritmo de busca por novas expressões ocorre no momento em que o algoritmo identifica uma árvore que atinja a quantidade de acerto do conjunto de validação determinada pelo usuário do sistema, ou quando o tempo máximo de execução do algoritmo é atingido. Vale ressaltar que o tempo máximo também deve ser definido pelo usuário.

O principal motivo da utilização do conjunto de validação para definição do critério de parada, evita que a técnica encontre uma árvore especialista apenas em um conjunto dos dados, condição indesejável quando o objetivo é a generalização.

4.3.6 Escolha da melhor Árvore a ser retornada

Após determinada a condição de parada do algoritmo de busca, uma árvore deve ser selecionada para fazer parte do conjunto de expressões do sistema de classificação, representando a classe pela qual a busca foi efetuada. Esta escolha é feita avaliando o conjunto de validação e o método de escolha consiste em localizar a árvore que obtiver o maior número de acertos e em casos de empate, aquela com menor valor de SQE.

5
Sistema de Apoio a Decisão (Sistema de Informação)

Está seção tem por objetivo apresentar o Sistema de Informação proposto para determinar o grau de malignidade de casos de câncer de boca. Serão apresentadas também algumas interfaces desenvolvidas para o sistema, com a finalidade de facilitar o entendimento deste.

Esta pesquisa apresenta um sistema inteligente de apoio à decisão utilizado para determinar o grau de malignidade de amostras de células de câncer de boca. Esses sistemas de apoio a decisão podem ser definidos de várias formas. Laudon *et al.* (2004) os definem como um conjunto de componentes inter-relacionados utilizados para coletar, processar, armazenar, recuperar e distribuir informação, com a finalidade de apoiar a solução de um determinado problema e proporcionar suporte à tomada de decisão. A utilização destes sistemas possibilita realizar determinadas tarefas com mais agilidade e precisão.

O sistema de informação apresentado por este trabalho foi elaborado utilizando a linguagem de programação JAVA e a estrutura desse sistema pode ser vista na Figura 28. Os componentes desta estrutura são:

- Base de dados de Amostra de Células cancerígenas;
- Núcleo do sistema;
- Base de conhecimento;
- Interface.

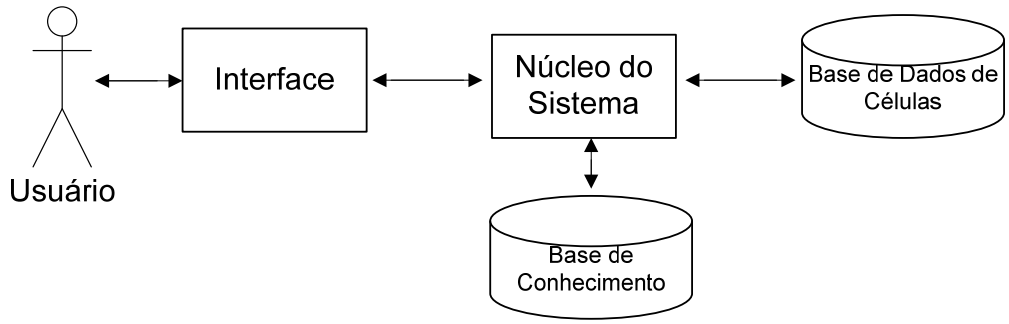

Figura 28. Estrutura do sistema de informação elaborado.

Todos estes componentes estão detalhados na Seção 5.1 até 5.4, a seguir.

5.1 Base de Dados de Células Cancerígenas

Esta base de dados consiste em armazenar dados de uma coleção de amostras de células cancerígenas, para a qual já é conhecida sua classificação. Ou seja, são coletados dados de várias lâminas e a cada uma delas deve ser informado também o grau de malignidade correspondente (a classificação). Vale ressaltar, que devem ser coletadas as mesmas características para todas as lâminas.

Esta base de dados será utilizada para a elaboração do conhecimento e é armazenada nesta pesquisa utilizando técnicas de serialização, ao invés de um Sistema de Gerenciamento de Banco de Dados (SGBD).

Segundo Sierra e Bates (2008, pág 260), "*A serialização permite simplesmente dizer 'salve este objeto e todas as suas variáveis de instância'*". É um método simples de ser utilizado em Java e sua escolha foi definida principalmente por:

- Poupar memória e processamento: Assumindo que para todas as execuções do algoritmo de aprendizagem é necessário recuperar os dados das células cancerígenas, sendo estes dados alterados poucas vezes, torna-se desnecessário realizar a mesma chamada ao banco de dados repetidamente.
- O sistema fica independente de qualquer tecnologia e utiliza menos espaço em disco: O objeto é guardado como um todo (atributos,

métodos e seus valores) sem precisar utilizar memória com MySQL, SQLServer, ou qualquer outro gerenciador de banco de dados.

Para a base de dados de lesões cancerígenas desenvolvida por esta pesquisa, foram selecionadas as seguintes características: quantidade de mitoses, queratinização, pleomorfismo, hipercromatina e relação núcleo/citoplasma.

5.2 Núcleo do sistema

O Núcleo do sistema consiste nas principais funcionalidades do sistema. É responsável por:

- Gerenciar perguntas e validar as respostas dos usuários, para a geração da base de dados;
- Gerar o conhecimento do sistema baseando nas informações obtidas pela base de dados. Para geração do conhecimento o sistema usa a técnica detalhada na Seção 4.3;
- Armazenar este conhecimento adquirido, para que este possa ser usado para futuras classificações;
- Classificar novos objetos que sejam descritos pelo usuário. Ou seja, o sistema deve receber algumas características do objeto a ser classificado e logo após deve processá-lo utilizando o conhecimento adquirido, e então, retornar a classificação.

5.3 Base de Conhecimento

Rezende *et al.* (2003) afirmam que "*A Base de Conhecimento contém a descrição do conhecimento necessário para a resolução do problema abordado na aplicação*". No caso desta pesquisa, o problema abordado é a classificacao de lesões malignas da boca e a base de conhecimento armazena o conjunto de fórmulas encontrados pelo algoritmo de aprendizagem para realizar a classificação.

O armazenamento desta base é feito através do mesmo método utilizado para armazenamento da base de dados de células, denominada serialização (Seção 5.1).

5.4 Interface Gráfica

A interface é a responsável pela transferência de conhecimento entre o usuário e o sistema em ambas as direções, ou seja, é responsável pela comunicação. Realiza a intermediação entre a representação interna do sistema e a representação mental do usuário (Rezende *et al.*, 2003). Devido a este fato, é imprescindível que está seja amigável e fácil de ser entendida pelo usuário.

As interfaces desenvolvidas para o sistema são divididas em categorias: interfaces principais, interfaces de gerenciamento de usuários, interfaces de gerenciamento de Bases de dados de células e interfaces de classificação. Estas podem ser vistas na Seção 5.4.1 até 5.4.4.

5.4.1 Interfaces Principais

A primeira interface a ser apresentada neste sistema consiste em campos utilizados para que o usuário informe seu usuário e senha, esta pode ser vista na Figura 29.

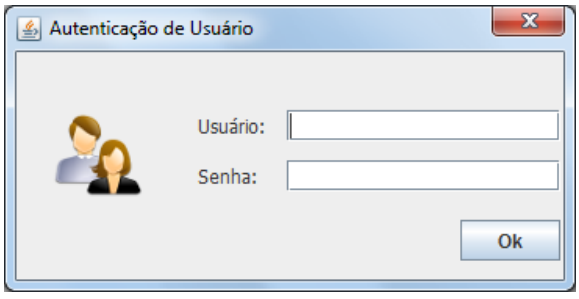

Figura 29. Tela de validação de usuário.

Logo após ter informado estes dados, o sistema os valida. Em caso de não consistência é exibido uma mensagem de erro, caso contrário, é apresentado a Tela principal do sistema. Esta pode ser visualizada na Figura 30.

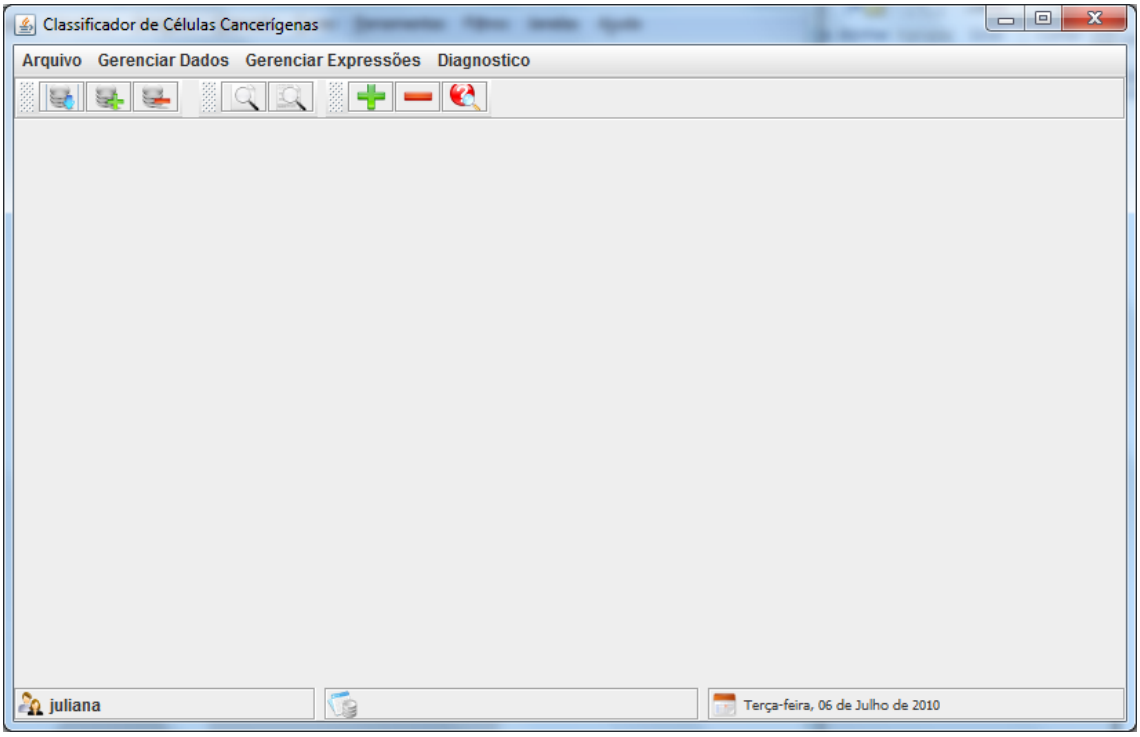

Figura 30. Tela Principal do sistema.

5.4.2 Interfaces de Gerenciamento de Usuários

Nesta categoria encontram-se as funcionalidades relacionadas ao gerenciamento de usuários: Adicionar Usuário, Excluir Usuário e Editar Usuários. Estas podem ser vistos na Figura 31. Duas outras características a serem observadas nesta figura, a opção "Sair" que finaliza o sistema e a possibilidade de "Troca de Usuário", nesta opção o sistema é bloqueado e logo após é exibido a tela de validação de usuário definida na Figura 29.

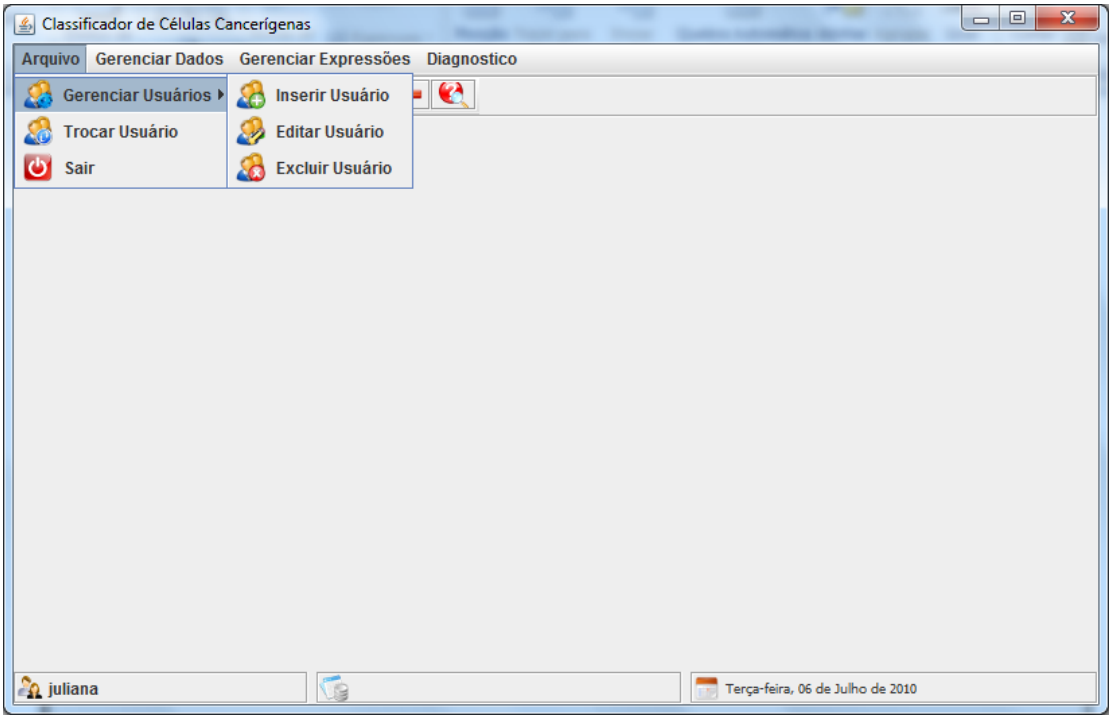

Figura 31. Tela de Gerenciamento de Usuários.

As interfaces utilizadas para adicionar um novo usuário, excluir usuário e editar os dados de um usuário são apresentadas abaixo na Figura 32, Figura 33 e Figura 34, respectivamente. Para adicionar, basta preencher os campos apresentados e confirmar clicando no botão "Adicionar". Já nas janelas de exclusão e edição, é necessário primeiramente clicar sobre o usuário que deseja efetuar a ação, e os campos de informação abaixo da tabela serão preenchidos com os dados deste usuário. Logo após, deve-se, no caso da exclusão, clicar sobre o botão "Excluir" e confirmar a ação, e no caso da edição, é preciso alterar a informação a ser modificada e logo após confirmar clicando no botão "Editar".

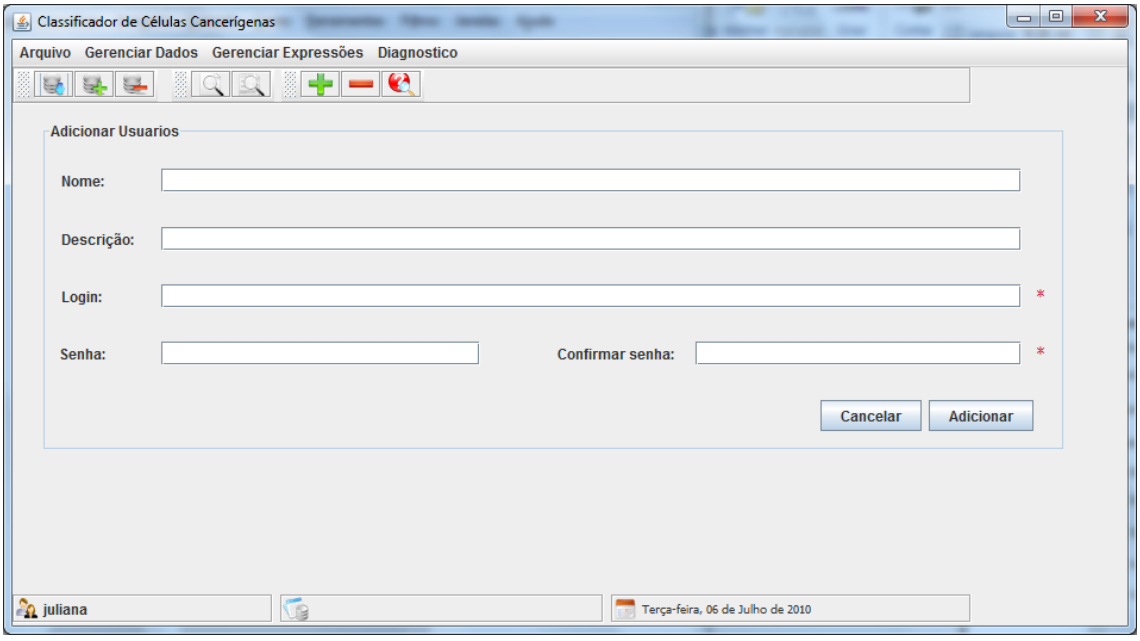

Figura 32. Inserir um novo usuário ao sistema.

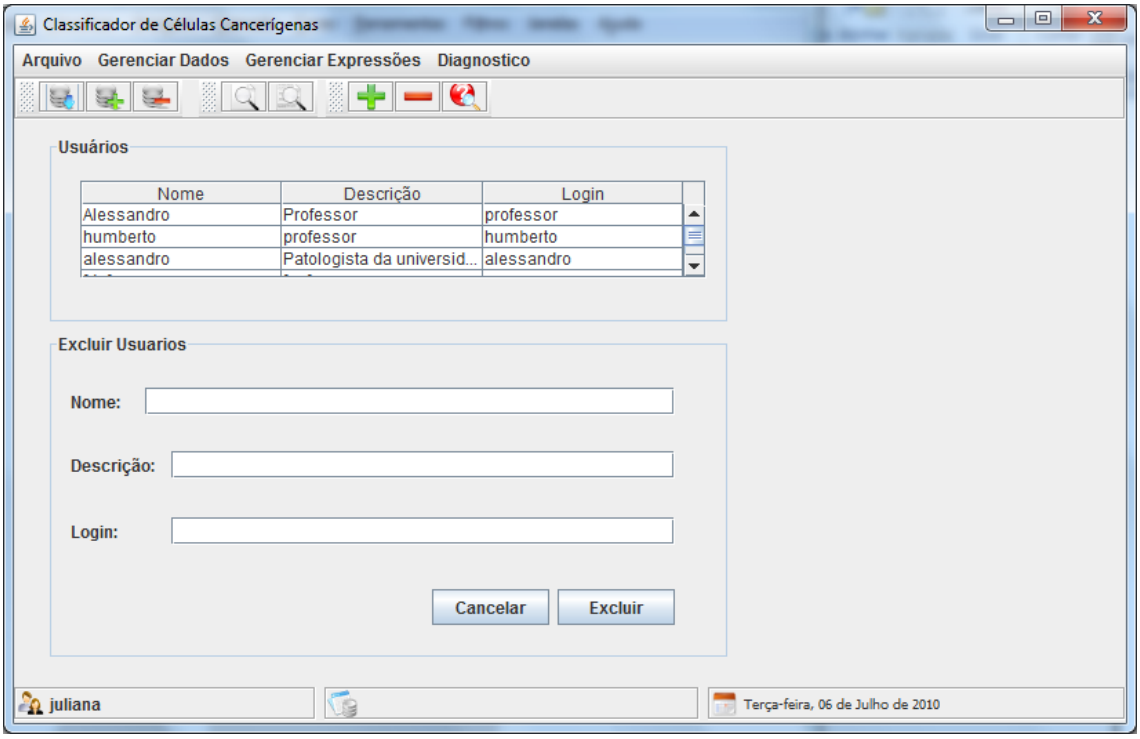

Figura 33. Excluir um usuário.

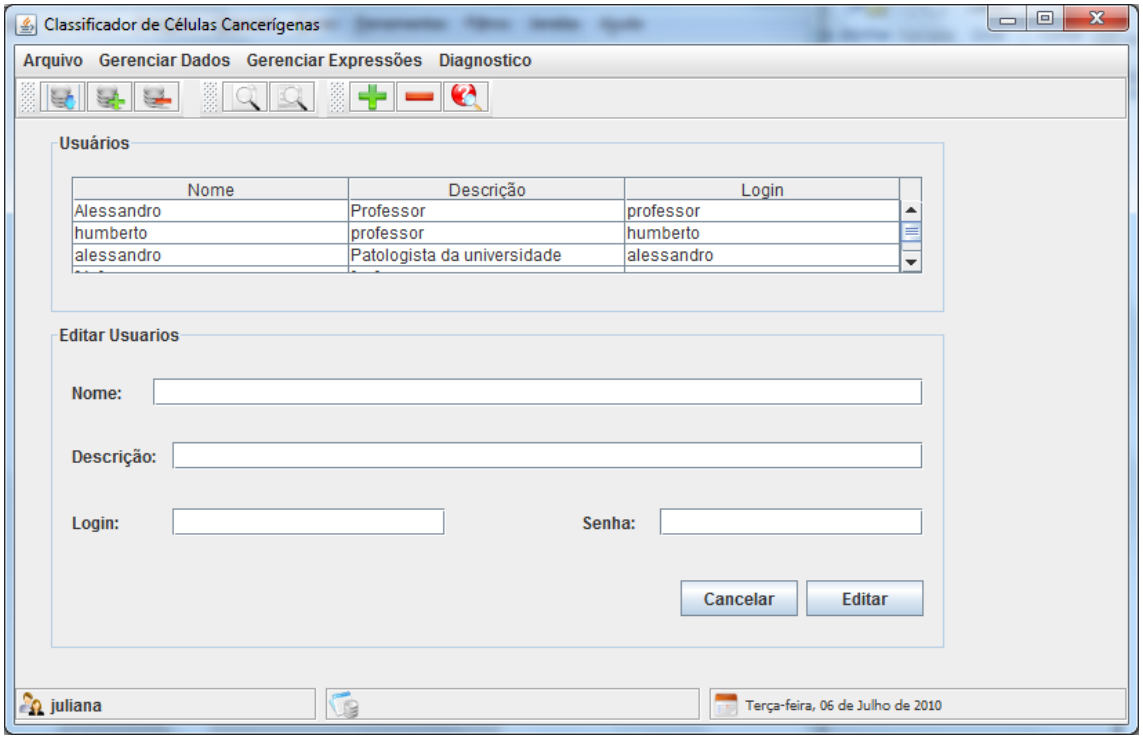

Figura 34. Editar dados de um usuário do sistema.

5.4.3 Interfaces de Gerenciamento de Bases de Dados de Células

O gerenciamento das bases de dados com informações das células de determinadas lesões, consiste em: adicionar e excluir base de dados.

Para adição de uma nova base de dados é preciso informar os dados dessa base e logo após escolher se os dados das células serão digitados manualmente pelo usuário ou serão lidos de um arquivo de texto. O formato deste arquivo pode ser encontrado no Apêndice A.

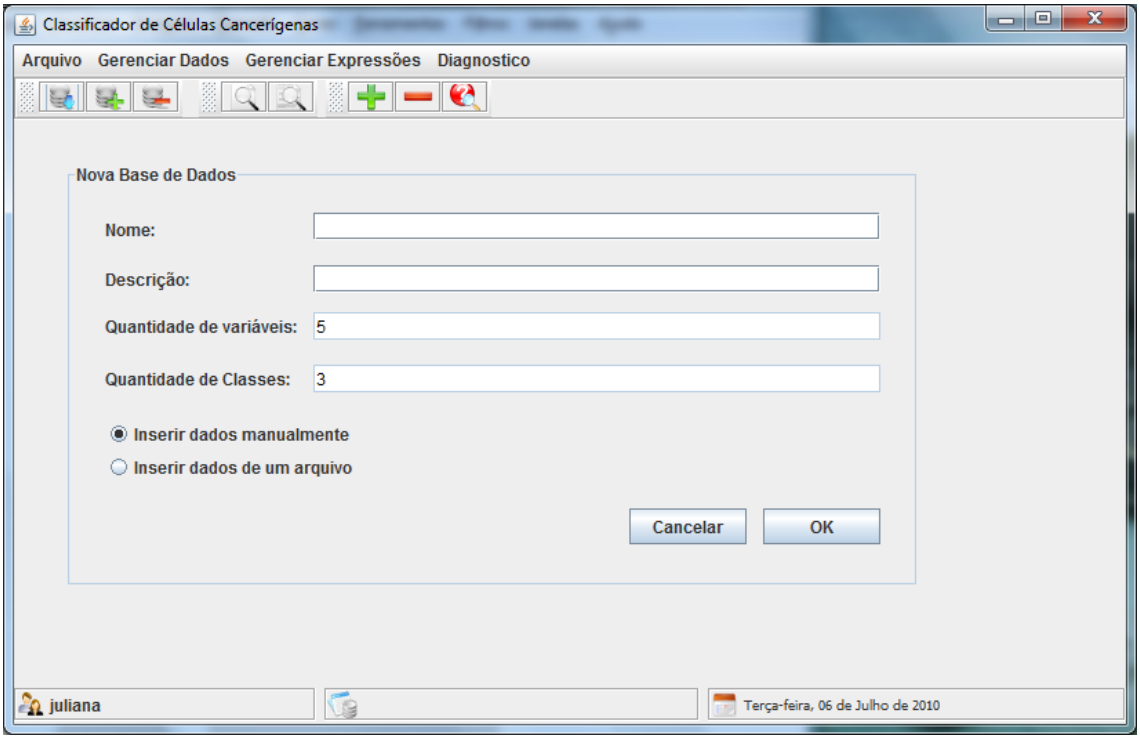

Figura 35. Adicionar Base de Dados.

Caso a opção escolhida seja "Inserir dados de um arquivo" a janela da Figura 36 é apresentada, e o usuário deve informar manualmente o caminho do arquivo ou procurá-lo através do botão "Buscar". Agora, se a opção escolhida for "Inserir dados de um arquivo", a janela da Figura 37 deve ser exibida e o usuário deve informar manualmente os dados de cada lesão, estes dados são: identificação da lâmina, o grau da lesão (Maligno I, Maligno II ou Maligno III) e as características das células de cada campo da lâmina.

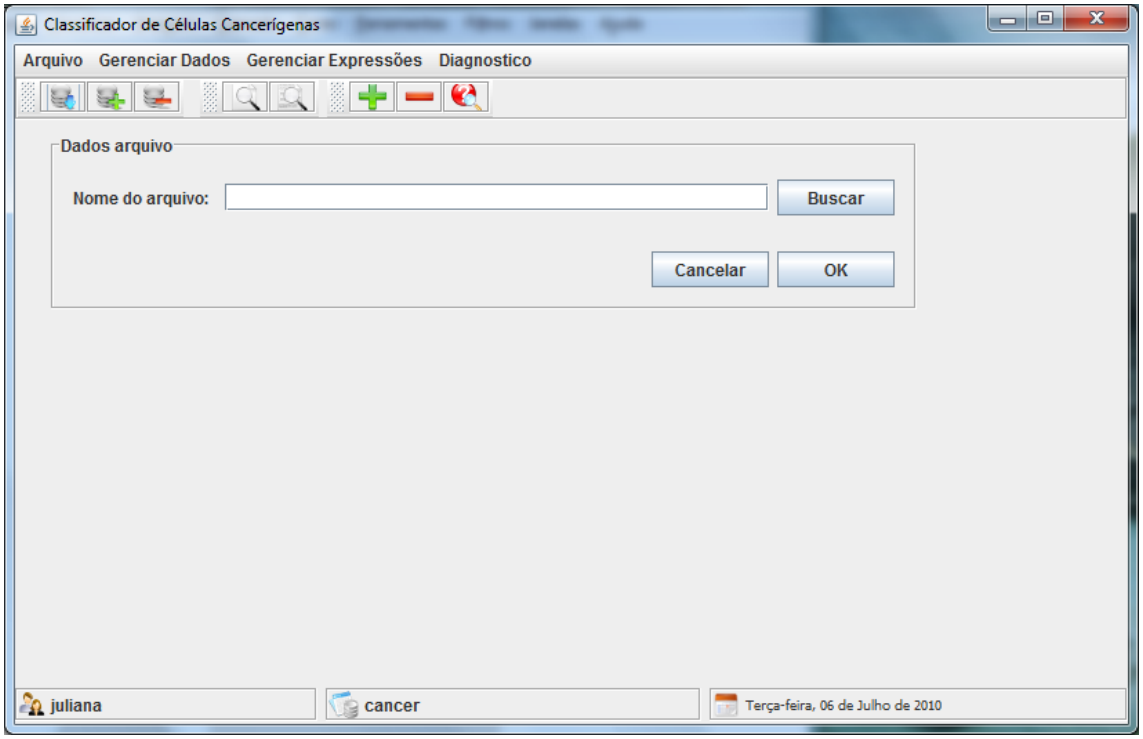

Figura 36. Inserir dados de um arquivo.

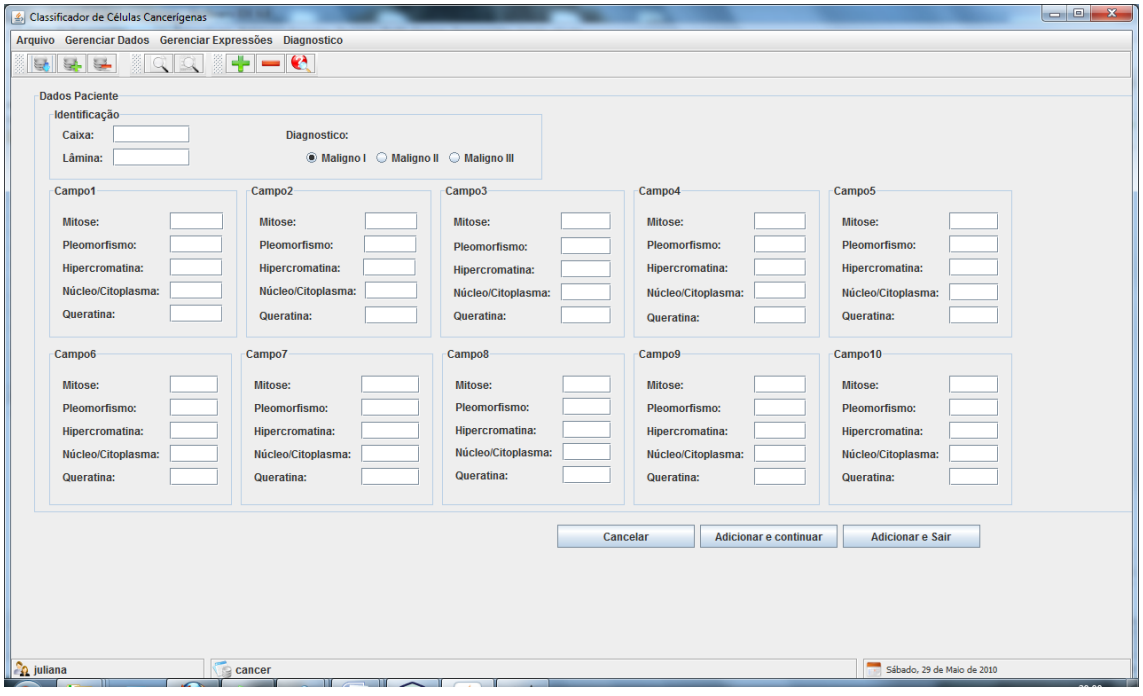

Figura 37. Inserir dados manualmente.

A exclusão de uma base de dados consiste em localizar esta base através da janela da Figura 38, logo após, será exibido uma mensagem para confirmação da exclusão. Sendo confirmado, a base será excluída.

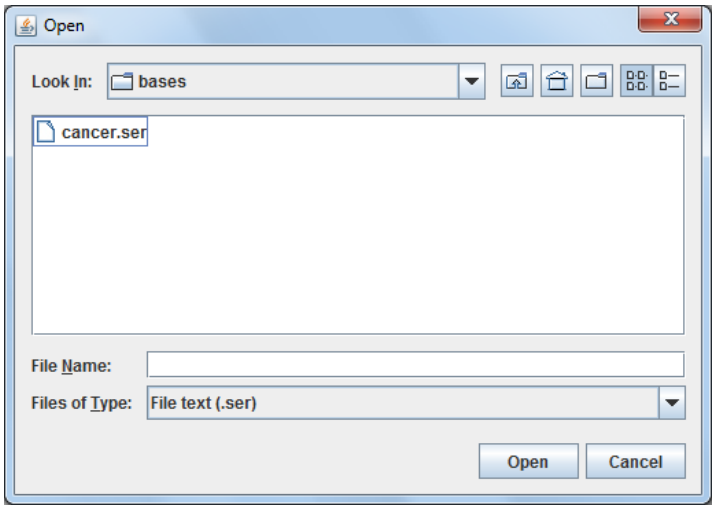

Figura 38. Localizar uma base de dados.

Agora as funcionalidades relacionadas aos dados das lesões de cada base de dados são: adicionar dados a uma base, excluir dados de uma base e visualizar dados de uma base.

Para adicionarmos dados de lesões a uma base já existente, usamos a janela representada pela Figura 37 ou a opção de leitura de arquivo da Figura 36.

Em casos de exclusão de lesões existentes em uma base (Figura 39), devemos abrir a base correspondente e logo após, selecionar o dado a ser excluído. Depois de selecionado, o sistema preenche os dados referentes a lesão no campo abaixo da tabela e deve-se confirmar a exclusão clicando no botão “OK”.

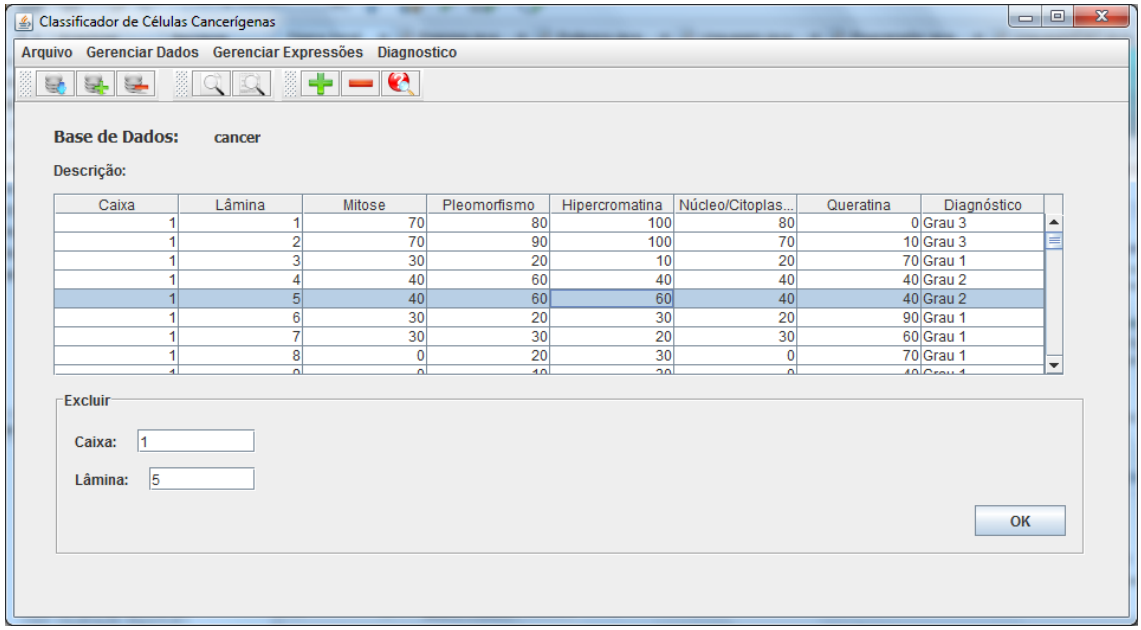

Figura 39. Excluir dados de uma lesão em uma base de dados.

A opção de visualizar os dados de uma base ocorre na tela referente à Figura 40, onde são exibidos dados descritivos da base selecionada e os dados de todas as lesões cadastradas.

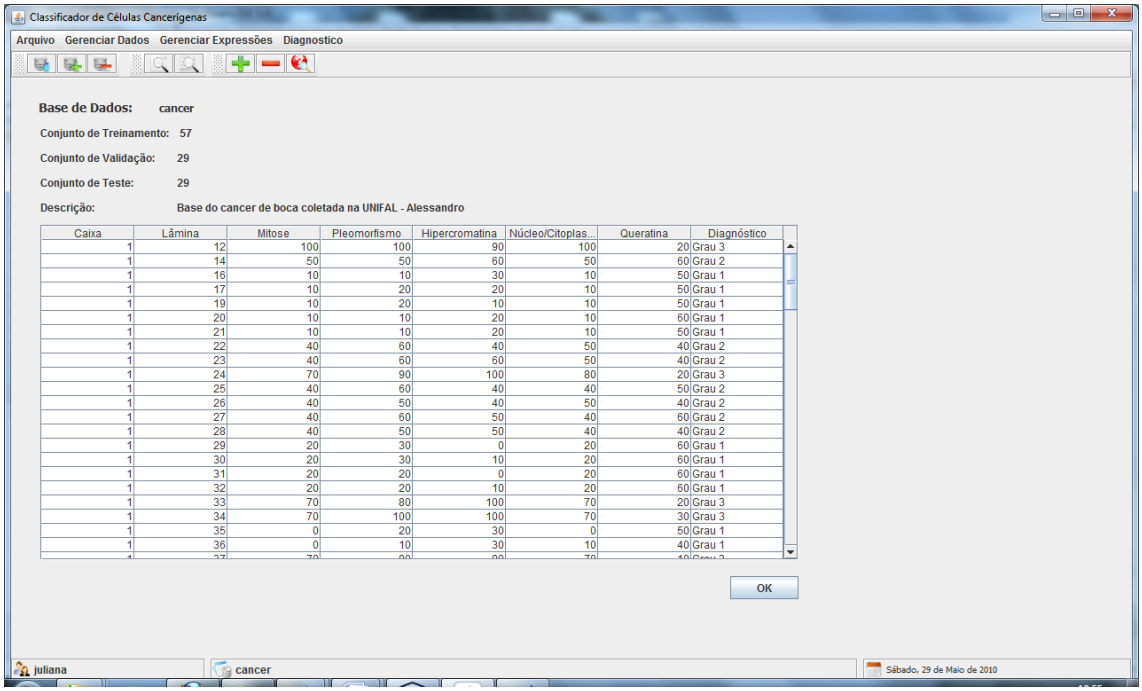

Figura 40. Visualizar dados de uma base.

5.4.4 Interfaces de Classificação

Esta categoria engloba três funcionalidades:

- Encontrar o conjunto de expressões responsável pela classificação do problema: O primeiro passo consiste em informar a base de dados a ser utilizada para geração dessas expressões, logo após, deve-se indicar qual forma de avaliação de expressões será utilizada: quantidade de acerto, valor do SQE ou ambas. Além disso, deve ser informada a quantidade de acerto a ser utilizadas pela condição de parada e o tamanho da população. Depois de gerada as expressões, cabe ao usuário aceitá-la ou não;

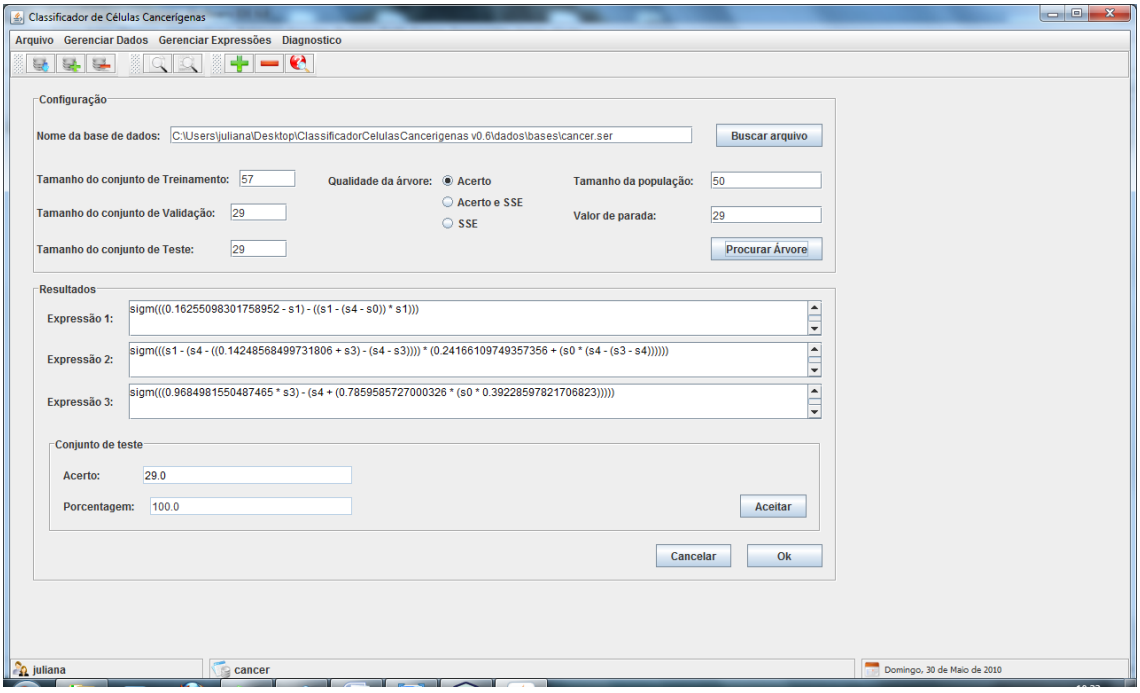

Figura 41. Tela utilizada para encontrar o conjunto de expressões para a classificação.

- Visualizar o conjunto de expressões atual: esta tela é utilizada apenas para exibir as fórmulas geradas para a classificação do problema;

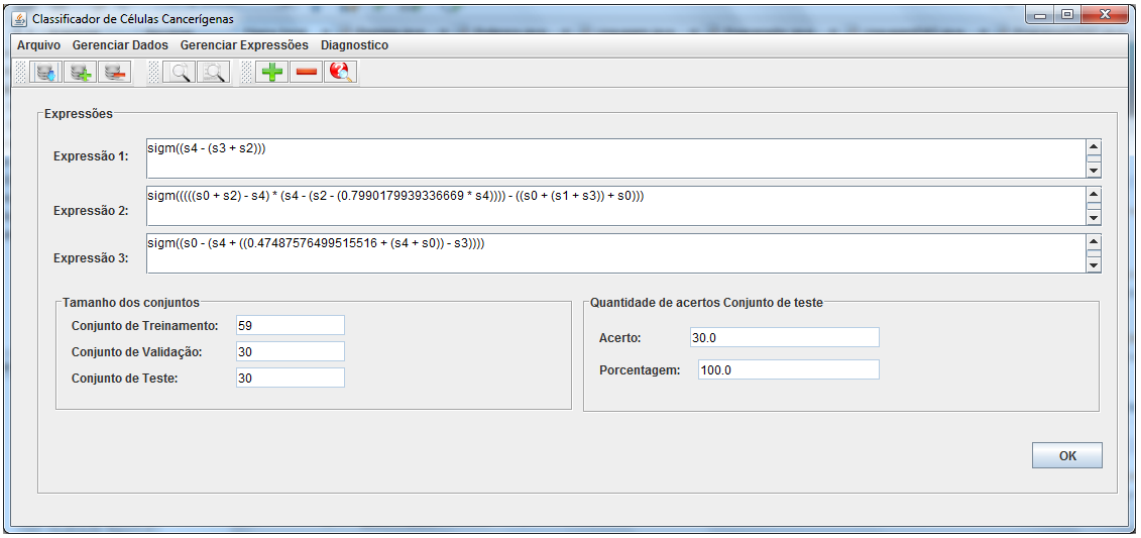

Figura 42. Interface de visualização do conjunto de expressões atual.

- Realizar a classificação de uma nova lesão: para realizar uma classificação, basta informar os dados da lesão e clicar no botão "Diagnóstico". No lugar do símbolo "*?*", apresentado na Figura 43, será exibido o resultado. Em casos que o sistema indique um resultado errado o especialista pode perfeitamente alterar este resultado selecionando a opção "Alterar diagnóstico" antes de salva-lo na base.

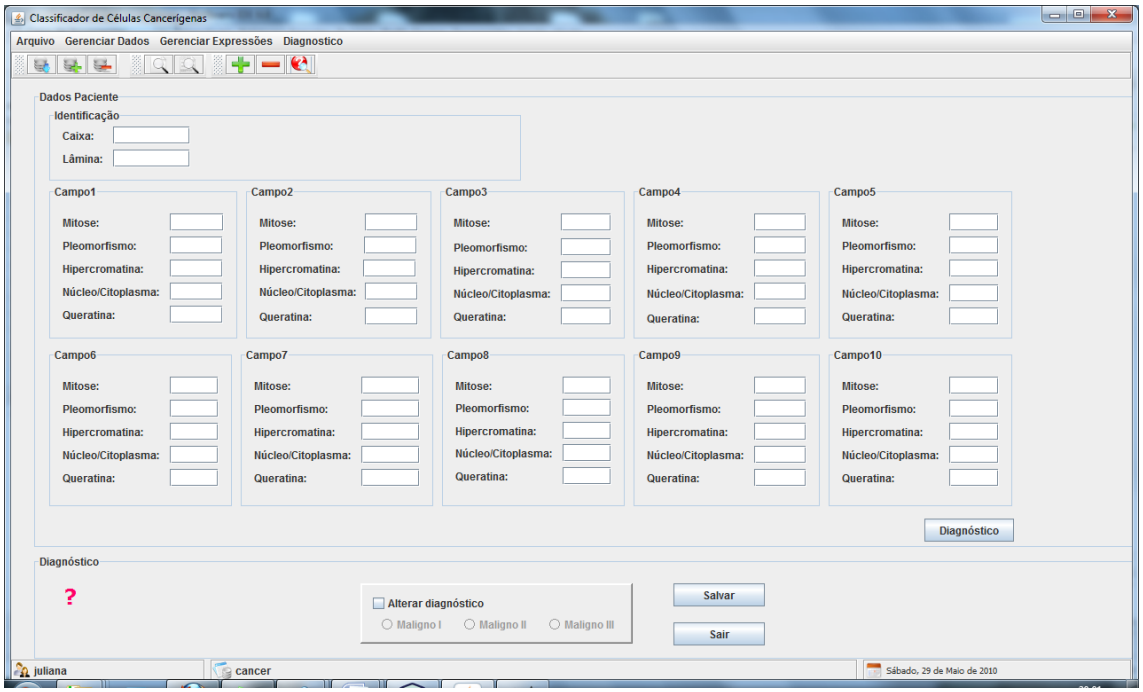

Figura 43. Interface utilizada para classificar novas lesões.

6
Resultados

Este capítulo é responsável por analisar e validar a técnica proposta. Divide-se da seguinte forma: a Seção 6.1 apresenta as bases de dados utilizadas para validação do sistema, na Seção 6.2, são apresentados os resultados alcançados e uma comparação com classificadores apresentados na literatura, e na Seção 6.3, são analisados os resultados obtidos.

O passo inicial do Sistema consiste na entrada com dados previamente conhecidos, ou seja, fornecer dados de algumas lesões cancerígenas informando as suas respectivas classes. O objetivo desta fase é o treinamento do Sistema Inteligente.

Logo após a localização das melhores expressões capazes de classificar o câncer através de características básicas da célula, o sistema estará pronto para realizar novas classificações (de células que ele desconhece). Neste ponto, basta informar os dados da lesão ao sistema e este retorna o diagnóstico.

Para realização dos testes experimentais e validação do sistema foram utilizadas 4 bases de dados detalhadas na Seção 6.1. Para cada experimento foram efetuadas 500 execuções e os detalhes de configuração do sistema e resultados são apresentadas na Seção 6.2.

6.1 Bases de Dados

O fato de utilizamos apenas uma base de dados para investigação do comportamento de um classificador, pode torná-lo tendencioso não possibilitando a avaliação da real generalização do sistema. Desta forma, esta pesquisa foi submetida a experimentos com 4 (quatro) bases de dados envolvendo problemas diversos. Essas bases são descritas na seção 6.1.1 e 6.1.2.

6.1.1 Base de dados Proben1

O Proben1 consiste em um repositório que fornece aos pesquisadores de sistemas inteligentes um conjunto de bases de dados para avaliação de desempenho dos sistemas desenvolvidos. Foi publicada no Relatório Técnico de Prechelt (1994) e pode ser encontrada no seguinte endereço eletrônico:

ftp://ftp.ira.uka.de/pub/neuron/proben1.tar.gz

Neste repositório são disponibilizadas 10 bases de dados para o problema de classificação, sendo escolhidas três para validação desta pesquisa. Esta escolha foi motivada pelo fato de querermos avaliar o desempenho obtido por classificadores binários (classificam em duas classes distintas) e classificadores ternários (classificação em três classes distintas). As bases binárias utilizadas foram: *Cancer* e Diabetes. E como exemplo de base ternária foi escolhido a base *Horse* e a desenvolvida pelo trabalho (detalhes na Seção 6.1.2).

Para cada uma das bases de dados do repositório, são disponibilizadas três bases distintas. Todas elas possuem os mesmos dados, porém, os padrões encontram-se em ordens diferentes, possibilitando assim, avaliar a real capacidade do classificador. Além disso, cada uma dessas bases é dividida em três conjuntos:

- Conjunto de treinamento (50% das amostras): utilizado durante o processo de otimização da Árvore Sintática, para aceitar ou não novas árvores obtidas;
- Conjunto de validação (25% das amostras): utilizado para parar o processo de otimização, ou seja, determina o momento de finalizar da busca por novas expressões;
- Conjunto de teste (25% das amostras): utilizado para avaliar a capacidade de generalização da Árvore Sintática adquirida.

Segue abaixo uma pequena descrição das bases utilizadas, sendo mais detalhes encontrados em Prechelt (1994).

- ***Cancer***: esta base de dados criada pelo hospital Norte-Americano Madison da Universidade de Wisconsin, descreve informações retiradas de células de câncer de mama reais, obtidas através de imagens digitalizadas. Para cada uma das 699 amostras da base, são

informadas 9 características das células, sendo estas, Espessura da massa celular, Uniformidade do tamanho da célula, Uniformidade do formato da célula, Adesão marginal, Tamanho de uma célula epitelial, Núcleo vazio, Cromatina branda, Nucléolo normal, Mitose. Além de indicar se a mesma possui câncer maligno ou benigno. Das 699 células, 65% possuem câncer benigno;

- ***Diabetes*:** apresenta informações pessoais de pacientes além de informações de exames médicos. Nesta base, existem dados de 768 pacientes, sendo que 500 destes pertencem à classe dos não-diabéticos (65,10%) e 268 à classe dos diabéticos (34,90%). Para cada uma das 768 amostras da base, são informadas 8 características dos pacientes, sendo estas: quantas gestações o paciente passou (Para o sexo masculino o valor atribuído é 0), concentração plasmática de glicose de 2h no teste oral de tolerância a glicose, pressão sanguínea diastólica, espessamento da prega cutânea do tríceps, insulina sérica de 2h, índice de massa corpórea (IMC), função de continuidade de diabetes e idade. Além de indicar se o mesmo possui diabetes ou não;
- ***Horse*:** esta base de dados é utilizada para prever o destino de cavalos com cólicas. São utilizados dados de exames veterinários para indicar se o cavalo vai sobreviver, morrer, ou se deve ser sacrificado. Existem dados de 364 cavalos, sendo que em 62% destes o cavalo sobreviveu, em 24% morreram e 14% tiveram que ser sacrificados. Para cada uma das 364 amostras da base, são informadas 58 características, entre elas: idade, tratado com alguma cirurgia ou não, freqüência respiratória, pulso, mucosa, entre outras. Além disso, indica se este morreu, teve que ser sacrificado ou sobreviveu.

6.1.2 Base de dados Câncer de Boca

Esta base de dados foi utilizada para treinar o sistema na classificação de casos de câncer de boca, informando o grau de malignidade destas lesões. As graduações possíveis são:

- Grau I
- Grau II
- Grau III

Foi elaborada no departamento de Patologia da Universidade Federal de Alfenas, pelo Professor Alessandro Antônio Costa Pereira. A seleção das características a serem utilizadas pelo sistema também foi elaborada por este professor, utilizando sua experiência no assunto e por meio de uma análise detalhada do problema.

As características são obtidas através de amostras de células da região lesionada, sendo estas utilizadas para geração de lâminas que serão analisadas por um microscópio. Vale ressaltar que essas lâminas são compostas por vários campos, e para a geração desta base foi definido que os dados consistem em uma média dos valores coletados de 10 destes campos. A escolha dos campos é realizada aleatoriamente.

A base elaborada contém informações de 135 casos, sendo 45 deles classificados em grau I, 45 em grau II e mais 45 em grau III. Para cada uma das 145 amostras da base, é indicado o rótulo da classe correspondente e são informadas 5 características das células da lesão, sendo estas: quantidade de mitoses, quantidade de queratinização, quantidade de pleomorfismo, relação núcleo/citoplasma e quantidade de hipercromatismo.

Foram geradas três bases de dados contendo as mesmas amostras, porém, em ordens diferentes. A base foi dividida em 3 conjuntos, assim como nas bases do repositório Proben1 (Seção 6.1.1).

6.2 Experimento

Para avaliar o desempenho da técnica proposta foram realizados experimentos sobre 4 bases de dados apresentadas na Seção 6.1. Na Tabela 1 apresentamos o nome de cada base, a quantidade de entradas para cada padrão e a quantidade de saídas do problema (quantidade de classes).

Tabela 1. Quantidade de características e de classes de cada base.

Nome	Quantidade de entradas	Quantidade de saídas
Cancer	*9*	*2*
Diabetes	*8*	*2*
Horse	*58*	*3*
CancerBoca	*5*	*3*

O sistema proposto é representado pela sigla ST (Sigm-Tree), sendo dividido em três configurações:

- ST1: O sistema classifica a melhor expressão pela quantidade de acerto;
- ST2: O sistema classifica a melhor expressão pela quantidade de acerto e o valor do SQE;
- ST3: O sistema classifica a melhor expressão pelo valor do SQE.

Para realização dos testes experimentais foi definido o tamanho da população igual a 50 expressões. Quanto ao critério de parada do algoritmo de busca, deve-se atingir 100% de acerto no conjunto de validação ou atingir o tempo máximo de 40 segundos de busca.

Os classificadores utilizados para comparação dos resultados foram obtidos em Tsakonas (2006) e Lin (2007), sendo eles:

- ES1 até ES2: (Programação Genética em Camadas) Classificadores baseado na Programação Genética Multi-populacional com diferenças paramétricas;
- DT: (Árvores de Decisão) As árvores de decisões são classificadores que representam uma tabela de decisão sob a forma de uma árvore. Gráfico com ramificações mostrando as combinações resultantes de várias combinações de condições;
- FRBS: (Sistemas *Fuzzy* Baseados em Regras) A Lógica *Fuzzy* tem como objetivo modelar o modo aproximado de raciocínio, tentando imitar a habilidade humana de tomar decisões racionais em um ambiente de incerteza e imprecisão. A idéia principal é que todas as coisas admitem graus;

- ANN: (*Perceptron* Multicamada com *Backpropagation*) As redes neurais artificiais são inspiradas no sistema nervoso biológico. São estruturas baseadas em ligações. Nós simples (neurônios) são interligados para formar uma rede de nós sendo esta estruturada baseada no cérebro humano.
- FPN: (Redes *Fuzzy-Petri* com Programação Genética) Sistema híbrido envolvendo redes *Petri* com lógica *Fuzzy* e Programação Genética.

Devido ao fato dos resultados destes classificadores serem obtidos de outras pesquisas, são realizados com diferentes quantidades de execuções, sendo a do sistema aqui descrito baseado em 500 execuções, como dito anteriormente, e a dos outros classificadores baseados em 10 execuções.

É importante ressaltar que o conjunto de validação determina a parada do algoritmo e o conjunto de teste avalia a generalização do método. Desta forma, os percentuais de acerto apresentados nas tabelas e gráficos são referentes ao conjunto de teste.

Na Seção 6.2.1 são apresentadas tabelas com a média, o desvio padrão e os melhores resultados obtidos por cada classificador, além de apresentar a quantidade de características dos padrões utilizadas para cada melhor fórmula encontrada no sistema aqui descrito. E na Seção 6.3 é realizada uma análise destes resultados.

6.2.1 Resultados

Os resultados obtidos para as bases de dados *Cancer*, *Diabetes*, *Horse* e *CancerBoca*, são apresentados na Tabela 2, Tabela 3, Tabela 4 e Tabela 5, respectivamente. Os dados informados pela tabela são:

- *Média* – Média dos resultados de todas as soluções encontradas.
- *DP* – Desvio Padrão dos resultados obtidos pelos experimentos.
- *Max* – Porcentagem de acerto da melhor solução encontrada.
- *C* – Quantidade de características dos padrões utilizada pela melhor solução.

Tabela 2. Resultados para a base de dados Cancer.

Classificador	Cancer1				Cancer2				Cancer3			
	Média	*DP*	*Max.*	*C*	*Média*	*DP*	*Max.*	*C*	*Média*	*DP*	*Max.*	*C*
ST1	98,16	0,62	**100**	9	94,69	0,80	96,55	9	96,17	0,76	97,71	7
ST2	**98,48**	0,59	**100**	6	95,05	0,68	96,55	7	95,98	0,70	98,28	9
ST3	97,83	0,54	99,43	9	**95,72**	0,60	97,13	9	95,90	0,50	97,71	9
ES1	97,7	0,72	98,85	-	94,89	0,69	96,55	-	96,32	0,78	97,13	-
ES2	97,7	0,54	98,85	-	94,6	0,48	95,4	-	96,09	0,71	97,13	-
ES3	97,82	0,85	99,43	-	94,89	0,92	95,98	-	96,38	0,61	97,13	-
ES4	97,76	0,79	98,85	-	94,94	0,59	95,98	-	**96,61**	0,57	97,13	-
ES5	97,7	0,77	98,28	-	94,77	0,74	95,98	-	96,32	0,40	97,13	-
ES6	97,82	0,80	98,85	-	94,83	0,47	95,4	-	96,03	0,69	97,13	-
DT	96,21	1.01	97,71	-	95,32	2,18	**98,28**	-	95,61	1,36	97,71	-
FRBS	95,61	1,42	97,71	-	95,55	1,23	**98,28**	-	95,1	0,83	96,56	-
ANN	94,34	1,24	97,13	-	91,7	2,16	97,13	-	94,72	1,7	**98,86**	-
FPN	95,69	0,94	97,13	-	95,17	1,19	97,71	-	95,58	1,43	97,71	-

Tabela 3. Resultados para a base de dados Diabetes.

Classificador	Diabetes1				Diabetes2				Diabetes3			
	Média	*DP*	*Max.*	*C*	*Média*	*DP*	*Max.*	*C*	*Média*	*DP*	*Max.*	*C*
ST1	74,64	2,08	**79,69**	8	72,54	1,50	**77,08**	8	76,61	1,68	**80,21**	8
ST2	**75,76**	1,18	79,17	6	73,14	1,43	**77,08**	8	77,32	1,14	**80,21**	7
ST3	75,35	1,22	78,65	8	73,96	1,29	**77,08**	8	**77,82**	0,90	**80,21**	8
ES1	72,50	2,76	78,13	-	71,25	2,44	75,52	-	75,16	2,53	77,60	-
ES2	72,71	2,04	75,00	-	71,46	1,80	75,00	-	75,99	2,72	78,65	-
ES3	73,91	2,2,4	77,60	-	71,46	1,32	73,96	-	75,36	1,32	78,65	-
ES4	73,13	1,98	77,04	-	71,88	1,15	73,96	-	76,16	2,13	78,65	-
ES5	72,08	1,94	75,52	-	72,29	2,19	75,00	-	75,16	1,04	77,08	-
ES6	71,98	2,44	76,56	-	72,08	1,63	74,48	-	75,47	1,91	77,60	-
DT	68,30	3,24	73,3	-	68,7	3,48	74,35	-	71,21	5,11	80,11	-
FRBS	73,53	3,40	78,02	-	**75,22**	1,22	76,44	-	75,75	1,64	78,01	-
ANN	75,46	1,26	77,49	-	74,59	1,15	76,97	-	71,24	1,84	75,92	-
FPN	73,18	2,56	76,97	-	72,92	2,65	76,97	-	71,79	2,16	75,91	-

Tabela 4. Resultados para a base de dados Horse.

Classificador	Horse1				Horse2				Horse3			
	Média	*DP*	*Max.*	*C*	*Média*	*DP*	*Max.*	*C*	*Média*	*DP*	*Max.*	*C*
ST1	65,65	4,14	76,93	51	61,97	3,54	71,43	50	61,10	3,95	70,33	51
ST2	**71,48**	3,54	**81,32**	47	**64,22**	2,57	70,33	33	**66,51**	3,07	73,63	38
ST3	68,60	3,57	79,12	58	63,74	3,10	**72,53**	58	62,31	3,81	**75,82**	58
ES1	64,51	4,61	71,43	-	61,98	3,78	67,03	-	61,21	3,11	65,93	-
ES2	66,59	3,28	73,63	-	61,98	2,15	64,84	-	62,31	2,69	68,13	-
ES3	64,73	1,90	68,13	-	62,31	3,81	70,33	-	62,42	2,73	68,13	-
ES4	65,05	3,39	71,43	-	62,31	2,98	67,03	-	61,43	2,95	63,74	-
ES5	65,38	2,65	68,13	-	61,65	3,61	68,13	-	63,52	3,18	67,03	-
ES6	65,38	4,22	73,63	-	63,63	3,38	68,13	-	61,65	0,96	62,64	-
DT	66,45	2,78	71,12	-	57,73	3,48	63,33	-	64,06	3,04	71,11	-
FRBS	58,52	4,49	61,12	-	59,73	4,29	62,23	-	64,72	5,00	72,23	-
ANN	69,39	2,05	71,12	-	56,50	2,55	61,12	-	63,73	2,53	68,89	-
FPN	59,63	3,90	63,33	-	61,12	4,84	64,45	-	58,89	4,44	63,33	-

Tabela 5. Resultados para a base de dados Câncer de Boca.

Classificador	CancerBoca1				CancerBoca2				CancerBoca3			
	Média	*DP*	*Max.*	*C*	*Média*	*DP*	*Max.*	*C*	*Média*	*DP*	*Max.*	*C*
ST1	98,49	2,49	100	4	98,06	2,71	100	4	99,47	1,25	100	4
ST2	**98,94**	2,23	100	4	98,68	2,32	100	4	99,59	1,17	100	4
ST3	98,8	2,06	100	4	**99,01**	2,14	100	4	**99,64**	1,45	100	5

6.3 Análise dos Experimentos

Observando a Tabela 2, Tabela 3, Tabela 4 e Tabela 5 foi concluído que os resultados obtidos pela pesquisa são promissores. Nas bases de dados *Cancer*, *Diabetes* e *Horse*, as quais foi realizada uma comparação com outros métodos, a pesquisa conseguiu alcançar ganhos se comparado aos classificadores encontrados em Tsakonas (2006) e Lin (2007). A Figura 44, apresenta os ganhos percentuais alcançados nas médias, comparando o melhor resultado alcançado pelos experimentos desta pesquisa com o melhor resultado dos outros trabalhos. Entre as nove bases, somente em duas não foi adquirido ganhos sobre os resultados da literatura.

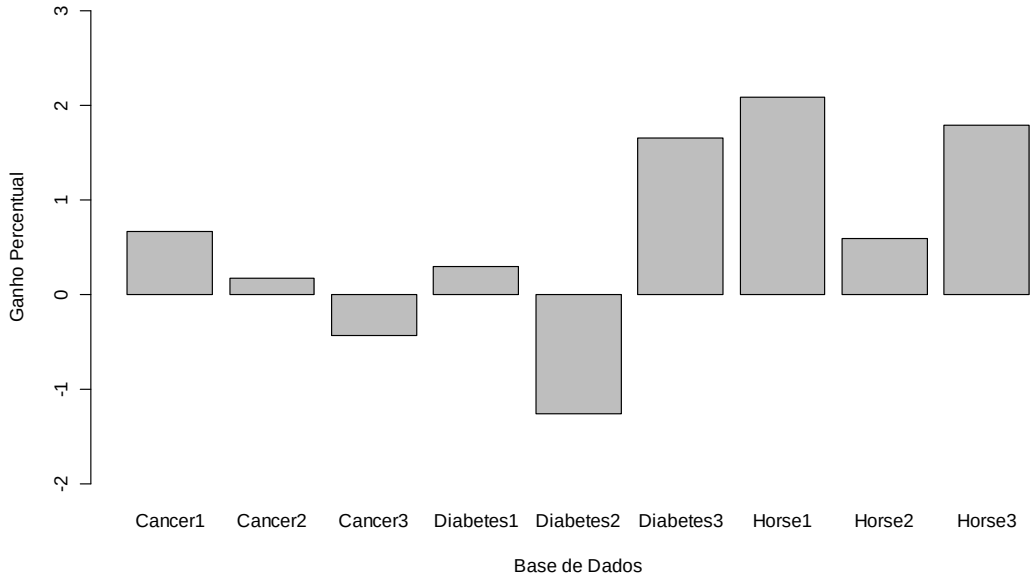

Figura 44. Ganhos percentuais obtidos pela pesquisa.

Outro fato importante a ser observado é a capacidade de selecionar apenas algumas características dos padrões para efetuar a classificação efetiva. Na tabela abaixo, são apresentadas as características de cada base que não foram utilizadas na melhor solução encontrada.

Tabela 6. Características não utilizadas na melhor árvore selecionada.

	ST1	ST2	ST3
Cancer1	-	3; **4**; **8**	-
Cancer2	-	**4**;**8**	-
Cancer3	**4**; **8**	-	-
Diabetes1	5;4	0;**3**	-
Diabetes2	-	**3**	-
Diabetes3	-	**3**	-
Horse1	0; 11; 25; **26**; 36; **41**; **55**	5; 10; 27; 38; **41**; **43**; 48; 51; **55**; 56; 57	-
Horse2	1; 28; **40**; **41**; **43**; 44; 50;52;	3; 4; 8; 9; **12**; 13; 14; 17; 19; 20; 21; **26**; 27; 28; 33; **40**; 42; **43**; 44; 48; 51; 53; **55**; 56; 57	-
Horse3	2; 6; **12**; 21; 32; **40**; **43**;	0; 1; 7; 10; **12**; 13; 16; 17; 19; 20; **26**; 27; 30; 31; 33; **41**; **43**; 47; 49; **55**	-
CancerBoca1	**2**	**2**	**2**
CancerBoca2	**2**	**2**	**2**
CancerBoca3	**2**	**2**	**-**

Analisando a Tabela 6 podemos observar que as características eliminadas para cada problema coincidem na maioria das bases correspondentes. Foram destacadas em negrito as que coincidem em pelo menos 3 configurações diferentes. Este fato pode reduzir consideravelmente o tempo e o custo na coleta de informações de futuras classificações.

Foram gerados também gráficos *boxplot* que possibilitam a analise da simetria e distribuição dos dados encontrados nos experimentos. Neles são apresentadas 5 características importantes para uma análise (Conceição, 2010):

- Mediana (q2): representado pela linha central da caixa conhecido também como segundo quartil. Representa o valor que divide a distribuição ordenada em duas partes iguais;
- Quartil inferior (q1) e Quartil superior (q3): representados pela parte inferior e superior da caixa respectivamente. O primeiro quartil é o valor que deixa um quarto dos dados abaixo e três quarto acima dele e o terceiro quartil é o valor que deixa três quartos dos dados abaixo e um quarto acima dele;
- As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o valor de q1 – 1.5*(q3 – q1) e do quartil superior até o valor de q3 + 1.5*(q3 – q1) representam o menor valor e o maior valor respectivamente;
- Os valores inferiores à haste inferior ou superiores à haste superior são representados individualmente no gráfico, sendo esses valores caracterizados como *outliers,* discrepantes ou valores atípicos (fora do intervalo de normalidade).

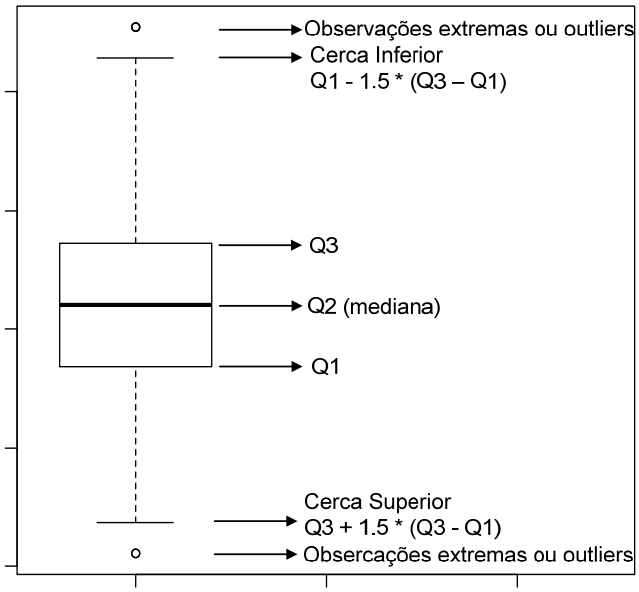

Figura 45. Principais características dos *Boxplot*.

Os gráficos *boxplot* para cada base de dados estão apresentados a seguir, onde o eixo das abscissas apresenta a configuração do algoritmo utilizada e no eixo das ordenadas temos a média do percentual de acerto do algoritmo testado. Para cada base, foram analisados os resultados sobre as 3 configurações diferentes.

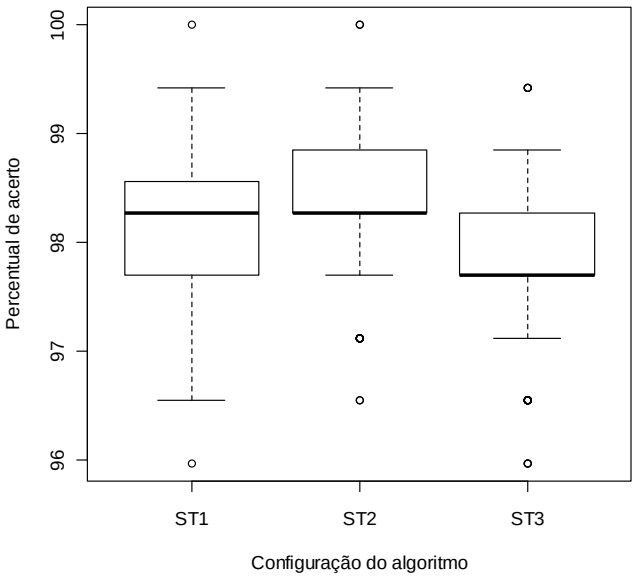

Figura 46. *Boxplot* referente a base de dados Cancer1.

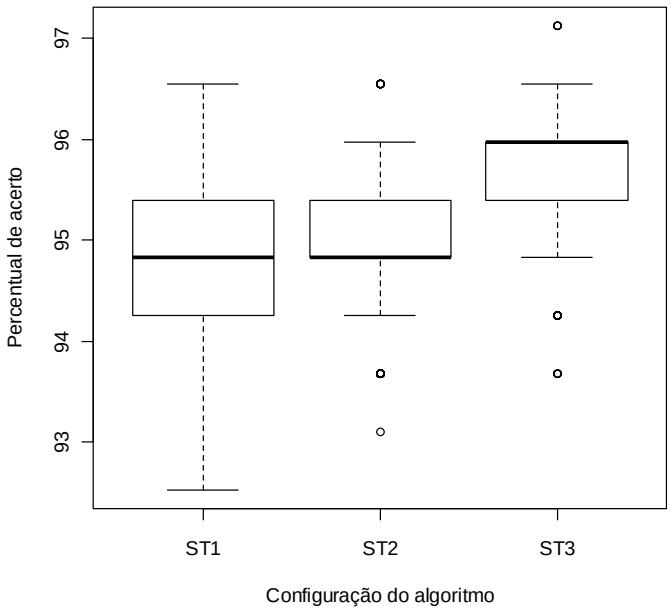

Figura 47. ***Boxplot*** **referente a base de dados Cancer2.**

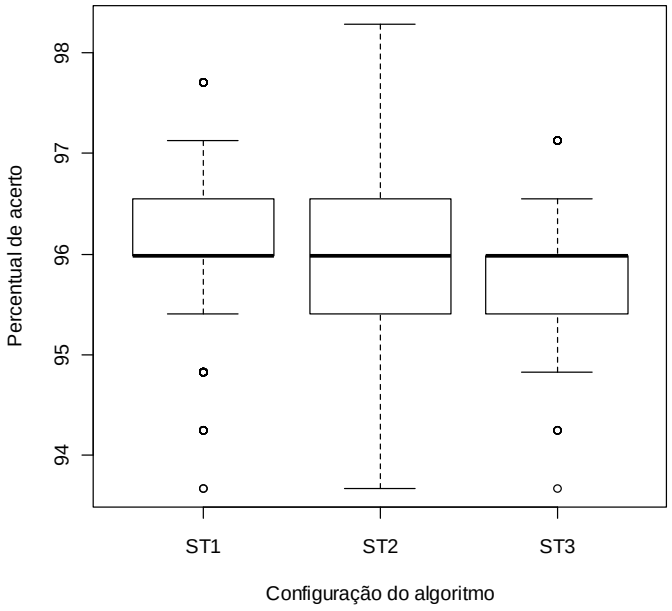

Figura 48. ***Boxplot*** **referente a base de dados Cancer3.**

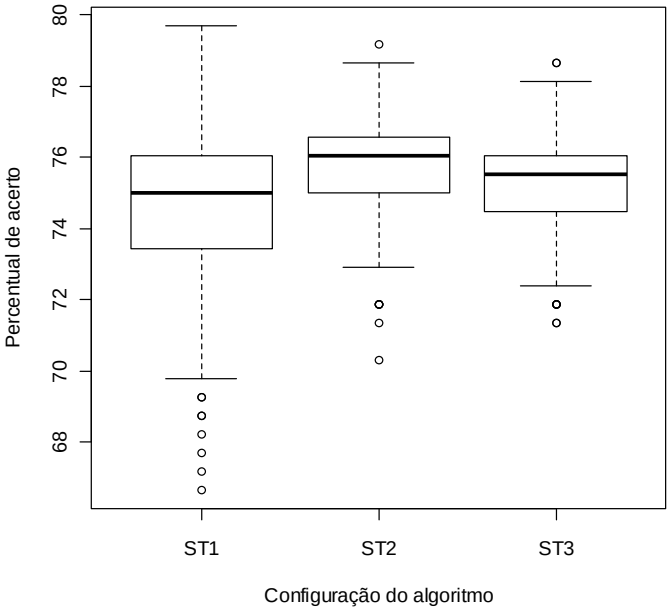

Figura 49. *Boxplot* referente a base de dados Diabetes1.

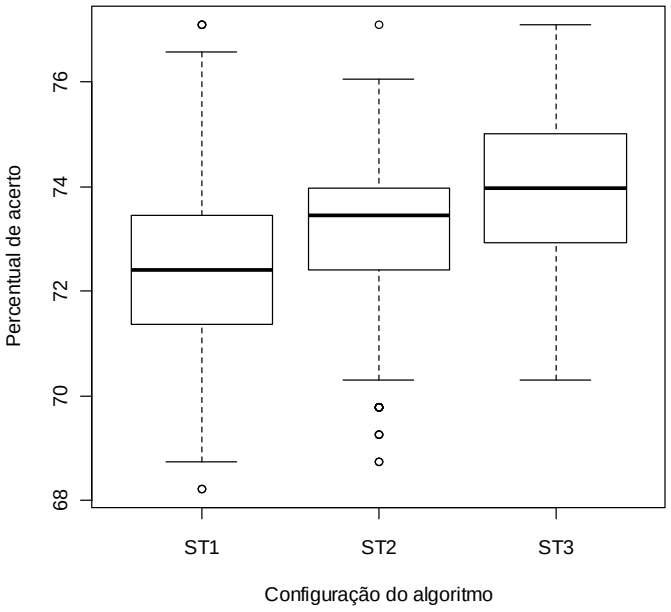

Figura 50. *Boxplot* referente a base de dados Diabetes2.

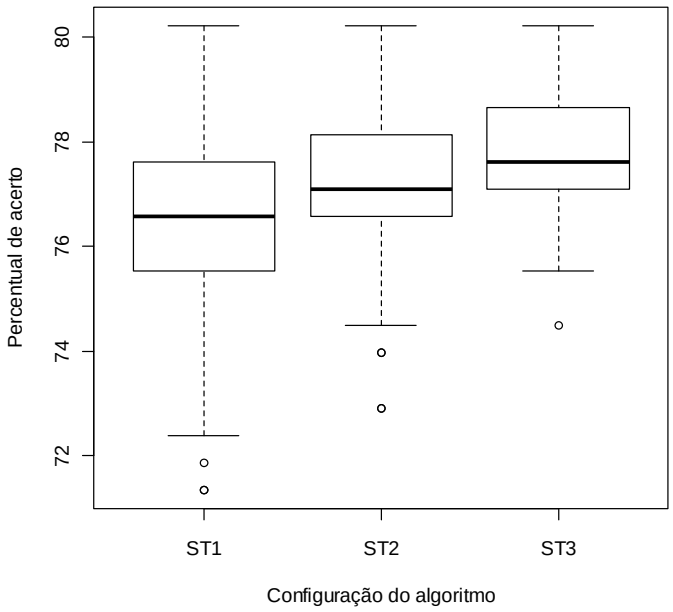

Figura 51. ***Boxplot*** **referente a base de dados Diabetes3.**

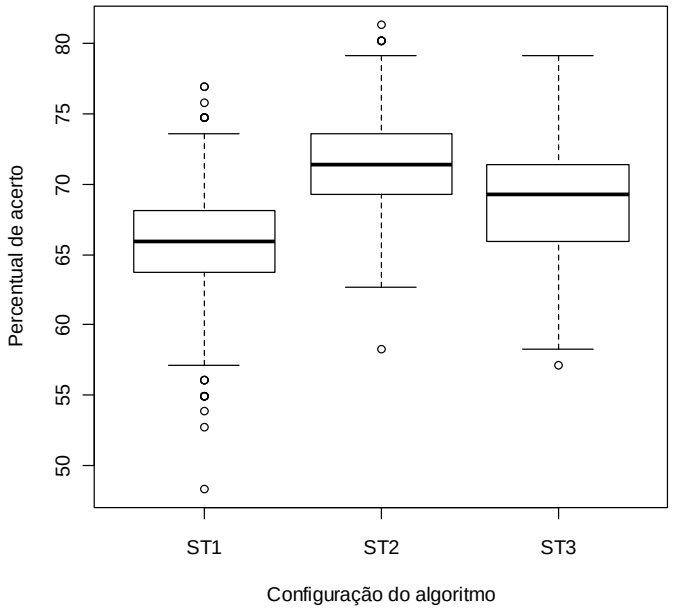

Figura 52. ***Boxplot*** **referente a base de dados Horse1.**

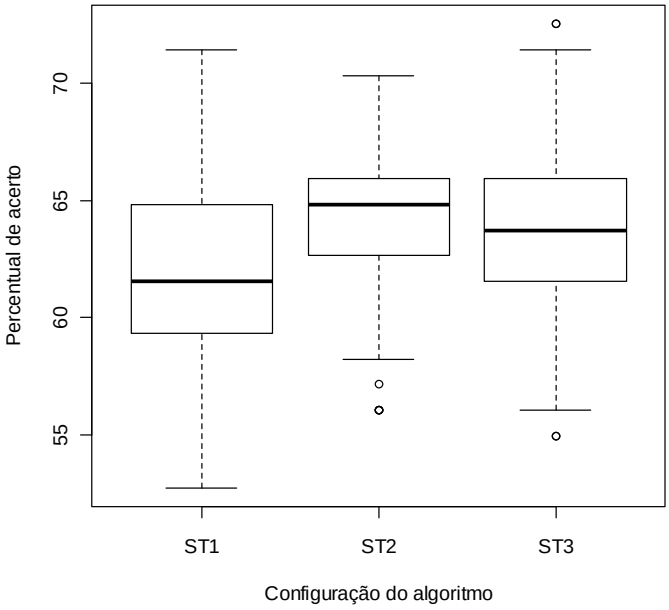

Figura 53. ***Boxplot*** **referente a base de dados Horse2.**

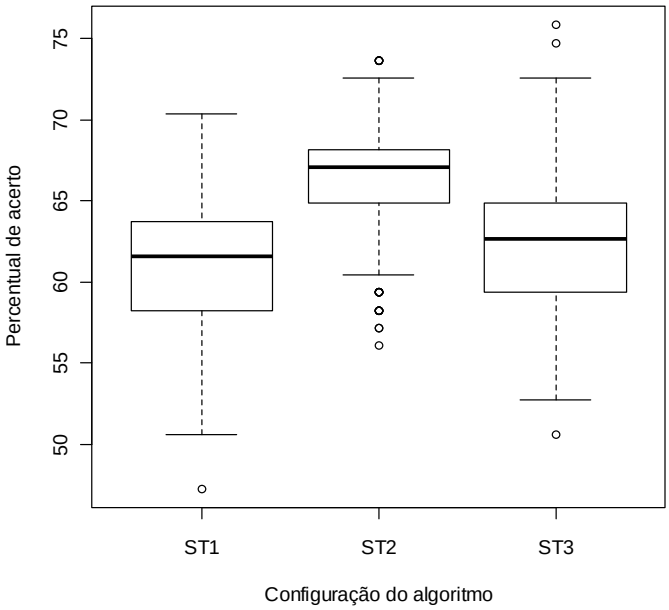

Figura 54. ***Boxplot*** **referente a base de dados Horse3.**

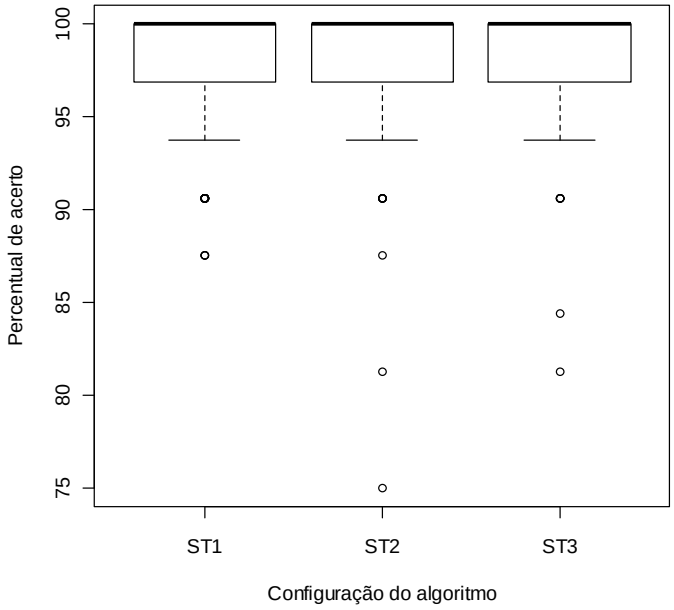

Figura 55. ***Boxplot*** **referente a base de dados CancerBoca1.**

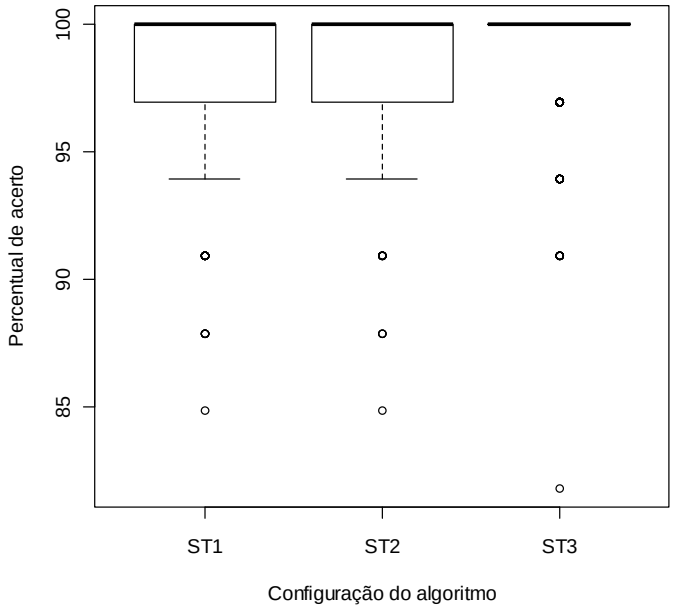

Figura 56. ***Boxplot*** **referente a base de dados CancerBoca2.**

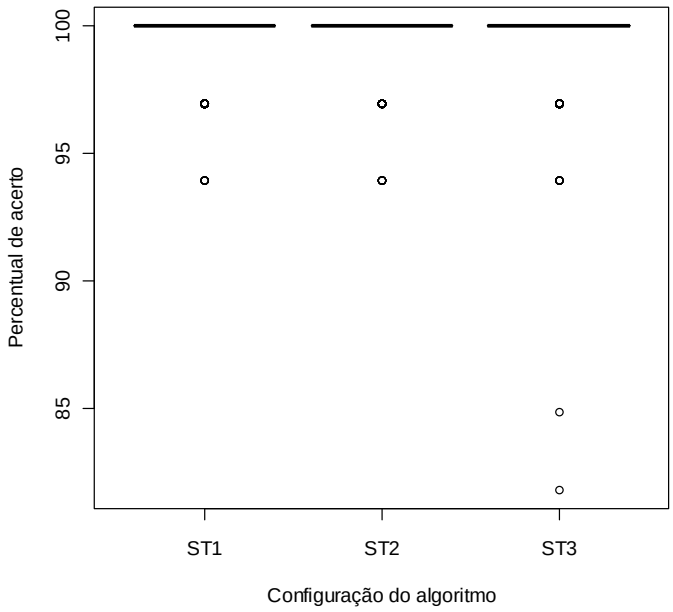

Figura 57. *Boxplot* referente a base de dados CancerBoca3.

Em relação a análise dos resultados pelos gráficos de *boxplot*, pode-se observar que o comportamento da configuração ST1 apresentou os resultados mais dispersos. Porém, através dos dados informados pelos gráficos não pode-se avaliar se todas as configurações produzem o mesmo efeito, tornando-se necessário uma análise estatística mais aprofundada em todos os conjuntos de configurações.

Arango (2005), afirma que os testes estatísticos baseados em valores, caso desta pesquisa, podem ser divididos em dois grandes grupos:

- Testes Paramétricos – se baseiam na hipótese de que as variáveis em análise apresentam uma distribuição normal;
- Testes Não-paramétricos – conhecidos também como testes de distribuição livre, neste não é necessário supor que a variável em análise apresenta distribuição normal.

Desta forma, antes de decidir o teste a ser realizado, foi investigada a distribuição das amostras a serem analisadas. O teste utilizado para essa verificação foi o Shapiro-Wilk. Esta escolha baseou-se na sua capacidade de adaptação a uma variada gama de problemas sobre avaliação de normalidade.

Os resultados apresentados pelo teste Shapiro-Wilk mostraram que as amostras não apresentavam normalidade, foi então utilizado para análise estatística, o teste de Kruskal-Wallis, conhecido também como Teste H. Este corresponde a um teste não-paramétrico e destina-se a casos de mais de duas opções na comparação de dados.

A hipótese nula deste teste é de que as distribuições das populações sob consideração sejam todas idênticas. Já a hipótese alternativa, é de que pelo menos uma das distribuições das populações seja diferente e que, por conseguinte, nem todas as distribuições das populações sejam idênticas (Mann, 2006).

As variáveis envolvidas neste teste foram: porcentagem de acerto e a configuração do sistema (ST1, ST2 e ST3). Analisando os resultados foi possível concluir que as configurações mais adaptadas aos problemas estudados foram a ST2 e ST3. Na Tabela 7 pode-se observar qual a configuração utilizada para o melhor resultado de cada base.

Tabela 7. Configuração utilizada para o melhor resultado de cada base.

Configuração	Quantidade de entradas
ST1	cancer3
ST2	cancer1, diabetes1, horse1, horse2, horse3 e cancerBoca1
ST3	cancer2, diabetes2, diabetes3, cancerBoca2 e cancerBoca3

A análise realizada possibilitou observar também, com 5% de significância, que as configurações não apresentam comportamento semelhante na maioria das vezes. Sendo encontrada semelhança apenas nas bases relacionadas ao problema do câncer de boca onde os resultados foram próximos do máximo.

7
Conclusões e Trabalhos Futuros

Este capítulo apresenta algumas discussões envolvidas nesta dissertação. É organizado da seguinte forma: Na Seção7.1, é apresentada de forma sucinta a proposta, as principais limitações do método, os resultados alcançados e as vantagens da técnica, e na Seção 7.2 são discutidas propostas para trabalhos futuros.

7.1 Conclusões

O objetivo central desta pesquisa foi o desenvolvimento de um sistema computacional inteligente capaz de contribuir com o patologista no diagnóstico correto do câncer de boca, auxiliando na escolha terapêutica, aumentando assim as chances de sucesso no tratamento do paciente.

O sistema determina o grau de malignidade de lesões cancerígenas e seu passo inicial consiste na entrada com dados previamente conhecidos, ou seja, fornecer dados de algumas lesões cancerígenas, informando as suas respectivas classes (grau de malignidade). Após receber estes dados o sistema passa por uma fase de aprendizagem que consiste basicamente em encontrar expressões matemáticas, que ao serem processadas com características numéricas das células cancerígenas, sejam capazes de determinar o grau de malignidade da lesão analisada.

Depois de localizada as melhores expressões, o sistema estará pronto para realizar novas classificações (de células que ele desconhece). Neste ponto, basta informar os dados da amostra ao sistema e este retorna o diagnóstico.

A técnica de aprendizagem mostrou-se adaptável para a realização de outros tipos de classificações, funcionando como uma forma de preservar, aproveitar e organizar o talento e a experiência de especialistas.

Testes foram realizados, e através dos resultados foi possível observar que o sistema pode ajudar consideravelmente os especialistas a realizar este diagnóstico. Em relação aos resultados comparados com os trabalhos Tsakonas (2006) e Lin (2007), foram obtidos ganhos percentuais sobre a média em 7 das 9 bases avaliadas. Estes ganhos são interessantes principalmente em casos de problemas que envolvem saúde pública. Por exemplo, nas bases relacionadas ao câncer, um pequeno percentual de acerto equivale a um ou mais diagnósticos realizados corretamente, aumentando as chances de sucesso no tratamento.

Alguns resultados podem ser destacados, como os obtidos pelas bases relacionadas ao câncer de boca, estes alcançaram médias entre 98,06% e 99,64%, o que representa mais de 32 acertos em 33 padrões de teste. Outro fator importante a ser destacado é que o maior desvio padrão foi 2,71%, que devido a quantidade de lesões utilizados para teste é pouco significante (menos de uma lesão). Todo o processo experimental pode ser encontrado na Seção 6.

Algumas vantagens importantes encontradas pela pesquisa são:

- Capacidade de selecionar automaticamente as características importantes para a classificação, diminuindo assim, o custo na coleta de informações antes do processo de futuras classificações; Um exemplo pode ser observado na base cancer1, apenas 6 das 9 características da base de teste foram utilizadas na melhor expressão do algoritmo ST2 que obteve uma alta taxa de precisão (100% de acerto no conjunto de testes);
- Possibilita ao patologista uma classificação mais objetiva e uniforme dos tumores analisados, minimizando as limitações de uso impostas pelos sistemas de classificação;
- Possibilita manter o conhecimento de profissionais experientes na organização e até mesmo o disponibilizar para ajudar patologistas menos experientes;
- Capacidade de armazenar o conhecimento de mais de um profissional, diminuindo o custo e o tempo gastos em casos que o patologista considere necessário a análise de mais de um especialista;

- Acelera o trabalho do profissional, sendo considerado apenas o tempo gasto na coleta dos dados. Com menos de um minuto de simulação é possível encontrar expressões que possuem taxa de acerto acima de 95% nas bases relacionadas ao câncer.

Uma limitação encontrada na utilização deste sistema é a necessidade de um especialista humano capaz de solucionar o problema em questão. Este especialista torna-se necessário, para a construção da base de dados que deve ser passada ao sistema durante o processo de aquisição de conhecimento. A elaboração desta base deve ser feita com cautela, pois se o número de exemplos for insuficiente, ou se os exemplos não forem bem escolhidos, o método de classificação encontrado pode ser de pouco valor.

7.2 Trabalhos Futuros

A presente dissertação apresentou um classificador de padrões utilizado para determinar o grau de desenvolvimento de lesões cancerígenas localizadas na boca.

Como dito na Seção 4.3, esta classificação ocorre da seguinte forma: para cada uma das classes do problema (Grau I, Grau II e Grau III) é localizado uma expressão correspondente. Após determinado essas expressões, são informados dados numéricos da lesão a ser analisada, sendo estes, processados pelas expressões. A que retornar o valor mais alto representa a classe escolhida.

Definindo aqui como classificador esse conjunto de expressões, ou seja, todas as fórmulas matemáticas utilizadas para representar as classes de um problema, uma melhoria a ser pensada é a geração de mais de um destes conjuntos. Desta forma, quando formos realizar a classificação, avaliamos a opinião de todos eles e a classe determinada mais vezes é a escolhida. Esta técnica funciona como se um paciente fosse buscar o diagnóstico com mais de um profissional. Para esta proposta pode ser utilizado também uma mistura de técnicas (RNA, PG, AG, etc.) sendo todas avaliadas na hora de realizarmos a categorização.

Em relação ao Sistema de Informação uma melhoria a ser alcançada consiste na automatização da leitura das lâminas, tornando assim o processo mais rápido.

8
Referências Bibliográficas

Abbas, A. K.; Kumar, V.; Fausto, N.; Mitchell, R. N. (2008). *Robbins Patologia Básica.* 8ª Edição. Rio de Janeiro: Elsevier.

Almeida, H. M. *Uma Abordagem de Componentes Combinados para a Geração de Funções de Ordenação usando Programação Genética.* Tese de Mestrado. Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

Anneroth, G.; Hansen, L. S. *A methodologic study of histologic classification and grading of malignancy in oral squamous cell carcinoma*. Scand J Dent Res. 1984;92:448-68.

Arango, H. G. (2005). Bioestatística – Teórica e Computacional. 2ª Edição. Rio de Janeiro: Guanabara KooGan S.A.

Bettiollo, L. *Aplicação de técnicas de Reconhecimento de Padrões para a investigação de Síndrome de Down no primeiro trimestre de gravidez.* Tese de Mestrado. Universidade Federal do Paraná, Curitiba, 2009.

Braga, A. P.; Carvalho, A. P. L. F.; Ludemir, T. B. (2000). *Redes Neurais Artificiais: teoria e aplicações.* Rio de Janeiro, RJ: Livros Técnicos e Científicos.

Brandwein-Gensler M. Teixeira M. S.; Lewis C. M.; Lee B.; Rolnitzky L.; Hille J. J.; Genden E.; Urken M. L.; Wang B.Y. *Oral squamous cell carcinoma. Histologic risk assessment, but not margin status, is strongly predictive of local disease-free and overall survival.* Am J Surg Pathol. 2005;29(2):167-78.

Broders, A. C. *The microscopic grading of cancer*. Surg Clin North Am. 1941;21(4):947-62.

Bryne, M.; Koppang H. S.; Lilleng R.; Stene T.; Bang G.; Dabelsteen E. *New malignancy grading is a better prognostic indicator than Broders' grading in oral squamous cell carcinomas*. J Oral Pathol Med. 1989;18:432-37.

Conceição, G. M. S.; Alencar, A. P.; Alencar, G. P. *Noções Básicas de Estatística*. Disponívem em: <*portal.saude.gov.br/portal/arquivos/pdf/apostila_estatistica.pdf*> acessado em 08 maio 2010. pág. 7,8. 2010.

Costa, A. L. L.; Pereira J. C.; Nunes A. A. F.; Arruda M. L. S. *Correlação entre a classificação TNM, gradação histológica e localização anatômica em carcinoma epidermóide oral*. Pesquisa Odontológica Brasileira;16(3):216-220. 2002.

Darwin, C. *The Origin of Species by Means of Natural Selection*. John Murray, London, 1859.

Eiben, A. E.; Smith, J. E. (2003). *Introduction to Evolutionary Computing. Natural Coputing*. Springer, 2003.

Ferreira, F. O. *O que é uma biópsia?*. Disponívem em : <*http://www.cco.med.br/index.php?option=com_content&task=view&id=21&Itemid=13*> Acessado em 10 maio. 2010.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. Macmillan Publishing Company.

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

Instituto Nacional de Câncer (INCA). 1996-2010. Apresenta informações gerais sobre o Câncer. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/boca/definicao>. Acesso em: 08 maio. 2010a.

Intituto Nacional de Câncer (INCA). *Atlas de Mortalidade por Câncer*. Disponível em: < http://mortalidade.inca.gov.br/prepararModelo00.action>. Acesso em: 08 Maio. 2010b.

Instituto Nacional de Câncer (INCA). *Estimativa 2010 : Incidência de Câncer no Brasil*. Rio de Janeiro: INCA; 2009.

Jakobsson, P. A.; Eneroth, C. M.; Killander, D.; Moberger, G.; Martensson, B. *Histologic classification and grading of malignancy in carcinoma of the larynx* (a pilot study). Acta Radiol Ther Phys Biol. 1973;12:1-8.

Koza, J. R.; *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

Koza, J. R. (1998) *Genetic Programming II, Automatic Discovery of Reusable Programs*. 2nd ed., MIT.

Laudon, K. C.; Laudon, J. P. (2004). *Sistemas de Informação Gerencias: Administrando a empresa digital*. 5ª ed. São Paulo. Prentice-Hall.

Lin, J.Y.; KE, H.R.; Chien, B.C.; Yang, W.P.(2007), *Designing a classifier by a layered multipopulation genetic programming approach*, Pattern Recognition 40 (2007) 2211-2225.

Lourenço, S. Q.; Schueler, A. F.; Camisasca, D. R.; Lindenblatt, R. d.; Bernardo, V. G. (2007). *Classificação Histopatológica para o Carcinoma de Células Escamosas*

da Cavidade Oral: Revisão de Sistemas Propostos. Revista Brasileira de Cancerologia , 53 (3), 325-333.

Mann, P. S. (2006). *Introdução à estatística*. 5ª Edição. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora S.A.

Minku, F. L.; Pozo, A. T. R; Vergilio, S. R. *Chameleon: uma ferramenta de programação Genética orientada a gramáticas*. Revista Eletrônica de Iniciação Científica, v. 3, n.2, 2003, ISSN 1519-8219.

Prechelt, L. *Proben1: A set of Neural Network Benchmark Problems and Benchmarking Rules*. Technical Report 21/94, Fakultät für Informatik, Universität Karlsruhe, 76128 Karlsruhe, Germany, September, 1994.

Rezende, S. O. *et al.* (2003). *Sistemas Inteligentes - Fundamentos e Aplicações.* 1ª Edição. Barueri, São Paulo, Brasil: Manole Ltda.

Rodrigues, E.; Lopes, H. S. *Inferência de gramáticas livres de contexto usando Programação Genética*. I Simpósio Brasileiro de Inteligência Computacional. Forianópolis, 2007.

Sierra, k.; Bates, B. (2008). *Certificação Sun® para programador Java™ 6 – Guia de Estudo (Exame 310-065)*. Rio de Janeiro, RJ, Brasil: Alta Books.

Silva, D. R. C. *Redes Neurais Artificiais no Ambiente de Redes Industriais Foundation Fieldbus Usando Blocos Padrões*. Tese de Mestrado. Universidade Federal do Rio Grande do Norte, Natal, 2005.

Souza, D. O. *Algoritmos Genéticos aplicados ao planejamento do transporte principal de madeira.* Tese de Mestrado. Universidade Federal do Parana, Curitiba, 2004.

Tsakonas, A. *A comparison of classification accuracy of four genetic programming evolved intelligent structures*. Inf. Sci. 176 (2006) 691-724. 2006.

Whigham, P. A. *Grammatical Bias for Evolutionary Learning*. Tese de Doutorado. Universidade de New South Wales, Austrália, 1996.

Wong, M. L.; Leung, K. S. (2000). *Data mining using grammar based genetic programming and applications.* Kluwer Academic Publishers.

9
Apêndice A

9.1 Formato do arquivo de entrada de dados do sistema.

Para entrada de dados no sistema através de arquivos texto, o seguinte padrão de arquivo deve ser respeitado:

- Não existe nenhum cabeçalho no arquivo;
- Cada linha representa uma lesão cancerígena;
- Os valores reais devem ser separados por ponto;
- Cada valor deve ser separado por espaço;
- Os dados devem estar na seguinte ordem: Mitose, Pleomorfismo, Hipercromatismo, Núcleo/citoplasma e Queratinização.

Um exemplo de arquivo pode ser visto na Figura 58.

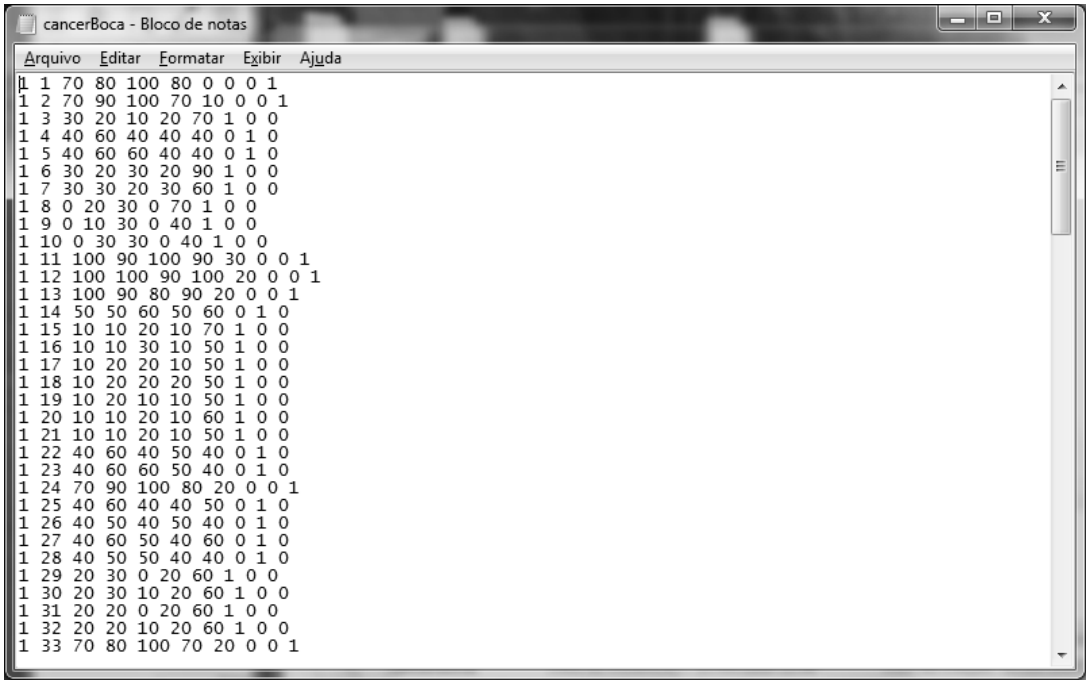

Figura 58. Modelo de arquivo de entrada de dados para o sistema.