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

Filipe de Oliveira Costa

MODELOS INTELIGENTES APLICADOS NA CLASSIFICAÇÃO DE PADRÕES DE CARACTERES

Alfenas, 02 de Julho de 2010.

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

MODELOS INTELIGENTES APLICADOS NA CLASSIFICAÇÃO DE PADRÕES DE CARACTERES

Filipe de Oliveira Costa

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. Dr. Ricardo Menezes Salgado

Alfenas, 02 de Julho de 2010.

Filipe de Oliveira Costa

MODELOS INTELIGENTES APLICADOS NA CLASSIFICAÇÃO DE PADRÕES DE CARACTERES

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. Flávio Barbieri Gonzaga

Universidade Federal de Alfenas

Prof. Humberto César Brandão de Oliveira

Universidade Federal de Alfenas

Prof. Ricardo Menezes Salgado (Orientador)

Universidade Federal de Alfenas

Alfenas, 02 de Julho de 2010.

Dedico este trabalho ao meu falecido tio Antônio Carlos de Oliveira, que neste exato momento está cumprindo a promessa que me fez há alguns anos.

AGRADECIMENTO

Agradeço A DEUS, por estar sempre me acompanhando durante todos os momentos de minha vida, me dando coragem, força e determinação, e não me deixando desistir de meus objetivos.

Aos meus pais Moisés e Maria José, e aos meus irmãos Mariana e André, que sempre estiveram ao meu lado me incentivando.

Ao professor Ricardo Menezes Salgado, por ser meu guia durante o desenvolvimento deste trabalho;

Ao professor Humberto César Brandão de Oliveira, por me orientar em minha iniciação científica, me ajudando sempre a corrigir meus erros;

A todos os demais professores da Ciência da Computação da UNIFAL-MG: Flávio, Tomás, Luiz Eduardo, Mariane, Melise, Paulo e Eliseu;

A todos os meus amigos, principalmente aos que conheci neste período de curso, por estarem sempre ao meu lado em momentos difíceis e também pelos momentos de distrações, alegrias, festas e bagunças.

E por último agradeço ao meu saudoso "irmão" e companheiro de república Danilo Fernandes Bispo, por sempre estar ao meu lado em momentos difíceis, pela sua amizade, pelos seus conselhos e, principalmente, por me mostrar que sempre é bom manter a cabeça fria.

Muito obrigado a todos.

"Algumas pessoas acham que o foco significa dizer sim para a coisa em que você vai se focar. Mas não é nada disso. Significa dizer não às centenas de outras boas ideias que existem. Você precisa selecionar cuidadosamente."

Steve Jobs

RESUMO

Nas últimas décadas houve um significativo avanço na tecnologia da informação. Formulários, cartas e outros documentos passaram a ser escritos com o auxílio de computadores e salvos em mídias eletrônicas. Mesmo com este avanço tecnológico ainda é possível encontrar uma quantidade exorbitante desses tipos de documentos em papel. Com o passar do tempo se tornou fundamental realizar o salvamento destes documentos em formato eletrônico, para que este possa ser salvo e editado em um computador, o que facilita seu manuseio e aumenta sua segurança. Porém a conversão destes documentos em material eletrônico ainda é muito custosa, visto que o que se faz atualmente é a re-escrita do documento no computador, muitas vezes feita por pessoas não qualificadas para o uso de *software* de edição de texto. Esta monografia de conclusão de curso apresenta propostas de modelos capazes de, através de uma imagem digitalizada de um documento, classificar os padrões de caracteres reconhecidos na imagem a fim de auxiliar no processo de digitalização do documento. Foram propostos dois modelos de classificação, o primeiro baseado no conceito de Sistemas Especialistas e o segundo modelo baseado em Máquinas de Vetores de Suporte *(Support Vector Machines* – SVM). Os modelos propostos obtiveram resultados satisfatórios no que diz respeito à quantidade de acertos e ao baixo custo computacional.

Palavras-Chave: Classificação de padrões, Classificação de caracteres, Sistemas especialistas, Máquina de Vetores de Suporte

ABSTRACT

In recent decades there has been a significant advance in information technology. Forms, letters and other documents have become written with the aid of computers and saved in electronic media. Even with this technological breakthrough is still possible to find an exorbitant amount of these types of documents in "paper". Over time, perform the rescue of these documents in electronic format so that it can be saved and edited on a computer became important, which facilitates handling and enhances its security. But the converting of these documents in electronic materials is still very costly, since what is done today is the re-writing of the document on the computer, often performed by people who are unqualified to use text editing software. This work presents proposals for models that may classify patterns characters on a scanned picture of a document to assist in the process of scanning the document. Two classification models were proposed, the first based on the concept of Expert Systems and the second model based on Support Vector Machines (SVMs). The proposed models obtained satisfactory results as regards the amount of hits and the low computational cost.

Keywords: Pattern classification, Characters classification, Expert Systems, Support Vector Machines

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

SVM	*Support Vector Machine* (Máquina de vetores de suporte)
SC	Sistemas de Classificação
RNA	Redes Neurais Artificiais
i.e.	"isto é"
OCR	*Optical Characters Recognition* (Reconhecimento Óptico de Caracteres)
BDT	*Bayesian Decision Theory* (Teoria Bayesiana da Decisão)
SE	Sistema Especialista
SBC	Sistema Baseado em Conhecimento
BC	Base de conhecimento
BT	Base de testes
PA	Pseudo-aleatória
Dimensão VC	Dimensão *Vapnik-Chervonenkis*
SVC	*Support Vector Classifier* (Classificador por Vetores de Suporte)

SUMÁRIO

1
Introdução

Este capítulo apresenta alguns detalhes sobre a confecção da monografia, bem como seus objetivos, justificativa e motivação para a realização deste trabalho de conclusão de curso.

Nas últimas décadas houve um significativo avanço na tecnologia da informação. Formulários, cartas e outros documentos passaram a ser escritos com o auxílio de computadores e salvos em mídias eletrônicas. Porém, ainda existem ambientes que possuem uma grande quantidade de documentos em papel. A tendência é realizar a conversão destes documentos em material eletrônico, porém esta conversão ainda é muito custosa, pois o que se faz atualmente é a re-escrita do documento no computador, muitas vezes feita por pessoas não qualificadas para o uso de software de edição de texto.

Uma possível solução para este problema é o desenvolvimento de um sistema capaz de "ler" o documento através de uma imagem digitalizada do mesmo e reescrevê-lo automaticamente, salvando em um formato digital, segundo Carvalho, Sampaio e Mongiovi (1999). Isto evitaria possíveis problemas causados pela re-escrita do documento por pessoas não qualificadas, além de facilitar o processo de gravação do documento em mídias digitais.

Considerando estes fatores, este trabalho de conclusão de curso foi desenvolvido com o intuito de propor modelos que possam ser aplicados na classificação de caracteres, visando auxiliar a digitalização de documentos.

1.1 Justificativa e Motivação

O desenvolvimento deste trabalho de conclusão de curso fez-se necessário para apresentar modelos computacionais que possam minimizar o problema da digitalização de documentos físicos, facilitando o processo de salvamento destes documentos em formato eletrônico.

Pode-se destacar como motivação os resultados esperados no início do desenvolvimento do projeto e a possível continuação deste trabalho em um programa de pós-graduação na área de inteligência artificial.

1.2 Problematização

Documentos importantes como formulários e contratos devem ser guardados de forma segura. Uma forma de aumentar a segurança destes documentos é salvando-os em mídias eletrônicas, como *CD, DVD, pendrive* e outros. A principal dificuldade é a digitalização destes documentos, que é realizada através da re-escrita do mesmo em software de edição de textos. Muitas pesquisas são voltadas para realizar a conversão a conversão destes documentos em materiais digitais automaticamente, i.e., sem que seja necessário realizar a re-escrita do mesmo. Existe um modo de auxiliar neste processo?

1.3 Objetivos

1.3.1 Gerais

O foco principal deste trabalho é propor modelos que realizem a classificação de padrões de caracteres isolados em documentos digitalizados, a fim de auxiliar no processo de conversão destes documentos em formato digital.

1.3.2 Específicos

Os objetivos específicos desta monografia são:

- Analisar algumas referências bibliográficas sobre classificação de padrões, com suas diferentes aplicações, destacando a técnica de classificação de caracteres;
- Descrever algumas propostas de classificação de padrões de caracteres aceitas na literatura, destacando suas vantagens e desvantagens;
- Descrever a base de dados utilizada nesse trabalho;
- Propor modelos de sistemas que sejam capazes de realizar a classificação de padrões de caracteres, com base no conjunto de dados utilizado;
- Apresentar os resultados obtidos, comparando-os com propostas bem aceitas na literatura.

1.4 Organização da Monografia

Esta monografia de conclusão de curso se encontra organizada da seguinte maneira: O capítulo 2 apresenta a descrição do problema de classificação de caracteres. No capítulo 3 são destacados os sistemas de classificação de padrões de caracteres e alguns trabalhos relacionados. Detalhes sobre a base de dados utilizada neste trabalho são apresentados no capítulo 4. O capítulo 5 apresenta a proposta da monografia. Os resultados obtidos por este trabalho são apresentados no capítulo 6. No capítulo 7 é feita a conclusão da monografia. Por fim, são apresentadas as referências bibliográficas utilizadas neste trabalho.

2
Descrição do problema

Este capítulo apresenta uma breve descrição sobre o problema de classificação de padrões de caracteres.

2.1 Considerações Iniciais

O avanço tecnológico das últimas décadas impactou positivamente na elaboração de textos. A escrita manuscrita de certos documentos se tornou rara. Livros, artigos, teses, documentos importantes como formulários, contratos e faturamentos, entre outros passaram a ser elaborados com o auxílio de um computador e um editor de textos digital. Com isto, houve uma redução considerável no tempo de elaboração de um documento e seu salvamento se tornou mais fácil, organizado e seguro, evitando o acúmulo de documentos em meio físico e a possível perda dos mesmos. Segundo Aires (2005, p.1), "documentos em papel parecem relíquias, principalmente quando se fala em manuscritos".

Por este avanço tecnológico ser recente, ainda existe uma quantidade exorbitante de documentos em papel. Para Oliveira *et al.* (2004), considerar que é rara a elaboração de documentos de forma manuscrita é um pré-julgamento falho, tendo em vista que a utilização do papel como meio de comunicação possui suas vantagens em relação a outros meios, como a padronização (não possui problema de interface com o escritor e o leitor), a portabilidade (apesar de ser mais lento que uma transferência eletrônica, seu transporte é bem estabelecido) e a não-necessidade de condições especiais para a escrita de um documento ou preenchimento de um formulário (são necessários apenas a habilidade do escritor, o papel e um instrumento se escrita). Porém a elaboração contínua de documentos em papel implica na necessidade de espaço físico disponível suficiente para comportar tamanha quantidade de documentos em determinados ambientes como

empresas e universidades, além de ser necessário um maior cuidado na organização e no armazenamento destes documentos. Além disto, dependendo da quantidade de material existente em um ambiente, a busca por um determinado documento no formato impresso ou manuscrito pode se tornar algo trabalhoso.

Uma alternativa para se resolver este problema é a digitalização do documento. Entretanto a conversão de um documento em conteúdo eletrônico atualmente é bastante complexa, visto que a técnica mais utilizada para a realização desta conversão é a re-escrita do documento, implicando em um gasto considerável de tempo para a realização desta conversão, que muitas vezes é efetuada por alguém não capacitado para utilizar ferramentas de edição de textos. Existem inúmeras pesquisas voltadas para a conversão de documentos físicos em material digital. A idéia principal destas pesquisas é desenvolver ferramentas que façam que o computador "leia" o teor impresso em um documento digitalizado através de um *scanner*, reconhecendo cada caractere apresentado na imagem digitalizada e transforme, de forma automática, o conteúdo obtido em um documento de texto digital dinâmico. Com base nestas informações, este capítulo irá apresentar uma descrição do problema do reconhecimento de caracteres, que é a base para um sistema de conversão de documentos físicos em material digital.

2.2 Reconhecimento de caracteres

O reconhecimento de caracteres é algo que vem sendo estudado amplamente pela comunidade científica desde a invenção do computador, segundo Veloso (1998). A elaboração de sistemas que façam este tipo de reconhecimento possui várias aplicações, como o desenvolvimento de leitoras automáticas de cheques bancários, máquinas automáticas de processamento de códigos postais, máquinas automáticas voltadas ao processamento de formulários preenchidos manualmente, entre outras (Aires, 2005).

A técnica de reconhecimento de caracteres se baseia na extração das características de caracteres de uma imagem de um documento digitalizado para que se possa fazer a classificação destas características obtidas. Para Aires (2005, p.3), "um fator determinante para um bom desempenho do reconhecimento é a seleção do conjunto de características a serem extraídas dos caracteres".

Existem alguns problemas que dificultam a realização do reconhecimento de caracteres. Pode ocorrer, por exemplo, de uma imagem digitalizada possuir baixa qualidade devido a algum inconveniente durante o processo de digitalização do documento, sendo necessário realizar um pré-processamento para a eliminação de ruídos na imagem. Outro problema que dificulta esta etapa é a existência de caracteres distorcidos, principalmente se tratando de documentos manuscritos, devido às características de caligrafia do escritor, que pode dificultar inclusive o reconhecimento dos caracteres por uma pessoa. Segundo Nunes (2004, p.5),

> a imensa capacidade humana de identificar, em uma determinada imagem, todos os caracteres ali presentes, acaba por permitir uma enorme quantidade de variações e imperfeições na escrita e conseqüentemente nas classes de símbolos ou padrões.

Pode-se destacar também como um possível empecilho o fato de um mesmo documento possuir vários tipos de caracteres diferentes como, por exemplo, letras maiúsculas e minúsculas, números, caracteres especiais, letras gregas utilizadas em fórmulas matemáticas, entre outros, o que implicaria no desenvolvimento de um sistema de reconhecimento de caracteres mais generalizado e complexo. Além disto, outra dificuldade a ser considerada é a semelhança entre alguns caracteres, como "I" e "J", "Q" e "O", "U" e "V", entre outros, que pode dificultar a classificação dos caracteres reconhecidos.

2.3 Considerações finais

O avanço da tecnologia influenciou na utilização do computador como meio de elaboração e salvamento de documentos de texto. Com isto, surgiram os sistemas de reconhecimento de caracteres, que atuam na conversão automática de documentos físicos em material eletrônico. Considerando estas informações, este trabalho de conclusão de curso visa apresentar um sistema que, com base em um conjunto de dados contendo padrões de caracteres reconhecidos, faça a classificação de suas características com base em um conjunto de regras. Com isto, pretende-se auxiliar no processo de digitalização de documentos físicos.

3
Revisão Bibliográfica

Este capítulo apresenta uma revisão bibliográfica sobre o tema abordado na monografia.

3.1 Considerações Iniciais

A classificação de caracteres é basicamente uma aplicação da técnica de classificação de padrões. Esta aplicação vem sendo estudado pela comunidade científica há algumas décadas. Várias propostas utilizando a classificação de padrões foram apresentadas com o intuito de facilitar a conversão de documentos físicos em material eletrônico, utilizando diversos modelos de classificação, além de diferentes conjuntos de dados.

Este capítulo apresenta uma breve descrição sobre os sistemas de classificação e suas aplicações, com foco na classificação de caracteres. Por fim, será apresentado o levantamento bibliográfico do tema.

3.2 Sistemas de classificação de padrões

Os sistemas de classificação de padrões são sistemas baseado em conjuntos de regras que são capazes de extrair informações de um ambiente e classificá-las (Santos, 2000). Geralmente estes sistemas são executados em duas etapas. A primeira etapa é o reconhecimento de padrões, que se baseia em extrair as características de padrões de um objeto ou evento desejado em um ambiente de busca, e a segunda etapa é a classificação de padrões, que tem como meta agrupar os padrões encontrados de acordo com suas características, a fim de minimizar a

variância de padrões intragrupos e, ao mesmo tempo, maximizar a variância de padrões intergrupos. Com estes grupos formados, é possível definir a qual deles um novo objeto ou evento encontrado irá pertencer. Esta classificação pode se tornar um processo complexo caso não haja a realização de uma busca sistemática visando encontrar simetrias existentes entre os padrões e que influenciam na eliminação de hipóteses possíveis gradativamente.

Os sistemas de classificação (SC) podem ser aplicados considerando duas vertentes: classificação de padrões supervisionada e classificação de padrões não-supervisionada. A classificação supervisionada pode ser descrita como processo onde as amostras de dados são conhecidas e classificadas antecipadamente e são utilizadas para classificar amostras de dados desconhecidas (Campbell, 1996 *apud* Máximo e Fernandes, 2005). Na classificação de padrões não-supervisionada ocorre o oposto, i.e., as amostras de dados não são previamente conhecidas e devem ser deduzidas a partir da base de conhecimento e agrupadas considerando a semelhança de característica entre padrões.

Existe uma gama de aplicações para os sistemas de classificação de padrões, que serão descritas a seguir.

3.2.1 Aplicações

Os SCs podem ser utilizados em vários ambientes distintos. Em Assis (2006) e Silva, Moita e Almeida (2008), foram desenvolvidos sistemas que têm por objetivo classificar e-mails recebidos por um destinatário qualquer em e-mail mal intencionado ou não. Este tipo de sistema visa aumentar a segurança em sistemas de e-mail, evitando a propagação de e-mails indesejados e vírus pela internet. Na área da saúde pode-se destacar o trabalho de Mancini *et al.* (2007), que apresenta um SC que trabalha com a classificação de padrões posturais de crianças respiradoras bucais ou nasais visando auxiliar o médico a definir um diagnóstico e/ou fazer a avaliação da evolução clínica de crianças com problemas respiratórios.

Também é possível aplicar SCs para a resolução de problemas na área da educação. O trabalho de Vieira *et al.* (2005) descreve uma ferramenta que monitora o diálogo entre alunos de ensino a distância, analisa as características particulares de diálogos entre alunos e professores de ensino a distância e classifica o ponto de vista de aprendizagem a fim de identificar os assuntos que despertam maior interesse dos alunos e os que causam mais dúvidas, ajudando a melhorar o processo de ensino/aprendizagem.

3.2.2 Reconhecimento e classificação de caracteres

Outra aplicação para os SCs é o reconhecimento e a classificação de caracteres em imagens. Esta técnica é denominada OCR (*Optical Character Recognition* – Reconhecimento Óptico de Caracteres). Segundo Rodrigues e Thomé (2000, p.1), a técnica OCR "estabeleceu as bases e a motivação para tornar o reconhecimento de padrões e a análise de imagens campos individuais de interesse da ciência".

Um sistema OCR é aplicado em quatro etapas. A primeira tem como objetivo a digitalização de um documento através de *scanners*, leitores ópticos, câmeras digitais, dentre outros. A segunda etapa é a etapa de reconhecimento dos caracteres. É nesta fase que o sistema realiza a extração de padrões dos caracteres da imagem digitalizada obtida na primeira etapa. Na terceira etapa é efetuada a classificação dos padrões dos caracteres obtidos na execução da segunda etapa. E por último é realizada a conversão da imagem obtida na primeira etapa em documento de texto em formato eletrônico, que possam ser editados a partir de software de edição de textos. A Figura 1 apresenta um modelo de um sistema OCR.

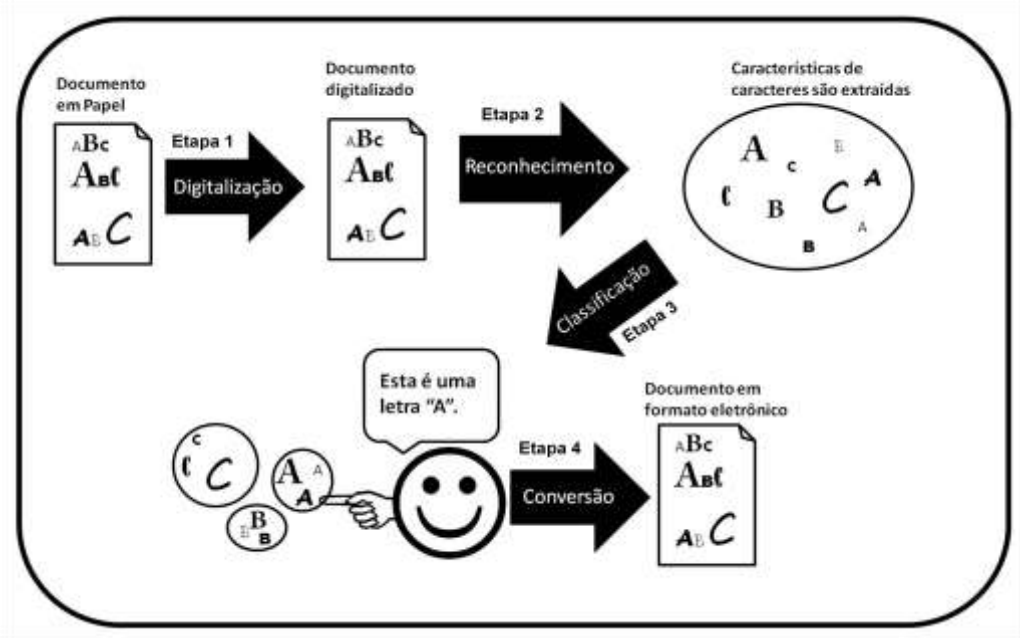

Figura 1: Modelo de um sistema OCR

Existem várias abordagens relacionadas aos sistemas OCR. Vários tipos de caracteres são avaliados, desde dígitos, letras, símbolos de outros idiomas, entre outros. A seguir serão apresentados alguns trabalhos relacionados sobre o tema, tratando a classificação de vários tipos de caracteres. Em seguida serão destacados três trabalhos que utilizam o mesmo conjunto de dados que este trabalho (a base de dados escolhida para este trabalho será destacada com detalhes no Capítulo 4).

3.3 Trabalhos relacionados

Serão apresentados neste tópico alguns trabalhos que realizam a classificação de caracteres.

O trabalho de Wang, Ding e Liu (2002) apresenta a proposta de um sistema baseado na teoria dos filtros de Gabor (Daugman, 1985) otimizado, que tem como objetivo realizar a extração de características de uma imagem para o posterior reconhecimento de caracteres do alfabeto chinês. Apesar da simplicidade do

método aplicado, o sistema se mostrou eficaz no que diz respeito à extração de características de caracteres para a realização do reconhecimento de padrões.

Batuwita e Bandara (2006) aplicam a técnica de lógica *Fuzzy* no reconhecimento de caracteres numéricos manuscritos. Inicialmente é realizada a fase de pré-processamento, onde a imagem é tratada até que se tenha a segmentação da imagem de cada caractere apresentado. Em seguida os segmentos obtidos de cada caractere são utilizados para calcular um conjunto de características *fuzzy*, que serão utilizadas para a realização do treinamento do sistema e para o reconhecimento dos caracteres obtidos. Segundo o autor, os resultados obtidos indicam que este método se mostra flexível para o reconhecimento de caracteres manuscritos de outros tipos, e não somente numéricos.

O trabalho de Moussa *et al.* (2009) é voltado para o reconhecimento de caracteres árabes utilizando uma técnica de análise de textura global. Este método é baseado em na geometria de fractais, e a extração de características independe do conteúdo do documento analisado. Primeiramente é obtida a distribuição da textura da imagem em duas dimensões. Feito isto, é aplicado uma técnica de classificação, uma vez que se torna possível diferenciar um tipo de letra do outro. Segundo o autor, o sistema poderia ser mais eficiente no que diz respeito à taxa de acertos obtida.

Já o trabalho de Ganapathy, Fernando e Davari (2005) mostra um sistema baseado em regras para o reconhecimento de caracteres alfanuméricos. De início é realizado um pré-processamento na imagem obtida. Em seguida são extraídas de cada caractere as características como linha horizontal, vertical, curvas, buracos entre outros. Com base nestas características é definido um conjunto de regras que será utilizado pelo sistema apresentado. A vantagem deste método é a ausência de treinamento. Porém não foram apresentados resultados de experimento, o que pode invalidar a confiabilidade do trabalho.

Em Velasques (2006) é apresentada uma técnica de classificação de pontos de segmentação de caracteres manuscritos. Após a realização da segmentação da imagem, o sistema primeiramente visa minimizar o número de hipóteses de segmentação para que possam ser posteriormente analisadas por um classificador de dígitos isolados. Esta minimização tem como objetivo filtrar hipóteses desnecessárias. Em seguida, com base nas hipóteses restantes é realizada a classificação dos caracteres através de um sistema de custos baseado na análise de curvas ROC (*ibidem*) e nas Maquinas de Vetor de suporte (*SVM – Support Vector Machine*). Os resultados apresentados mostram que existe uma redução do custo computacional devido à eliminação de hipóteses desnecessárias. Além disto, foi obtida uma taxa de acerto na classificação de caracteres relativamente alta.

3.3.1 Trabalhos relacionados que utilizam o mesmo conjunto de dados

A seguir será apresentada uma breve descrição dos trabalhos de Frey e Slate (1991), Schwenk e Bengio (2003) e Husnain e Naweed (2009), os quais foram utilizados na realização de comparação dos resultados obtidos por este trabalho pelo fato de trabalharem com o mesmo conjunto de dados.

3.3.1.1 Frey e Slate (1991)

A proposta do trabalho de Frey e Slate (1991) é definida como um classificador de padrões supervisionado voltado para o reconhecimento de caracteres do alfabeto inglês (A-Z). Este classificador é baseado na técnica de classificação adaptativa Holland. Trata-se de um classificador adaptativo baseado em regras e voltado para a classificação de padrões em geral, que se baseia em gerar um conjunto de regras e realizar a classificação de padrões com base neste conjunto gerado. A criação de novas regras pode ser efetuada por uma grande variedade de métodos, como geração aleatória de regras, utilização de algoritmos evolutivos, generalização e

especialização, entre outros. As regras geradas são analisadas e as menos importantes ou inválidas são descartadas. Ao final da execução destas técnicas se obtém um conjunto de regras que formará a base do sistema de classificação. A classificação adaptativa Holland, bem como suas respectivas técnicas de geração de regras, é descrita de forma mais detalhada em Holland *et al.* (1986).

No classificador proposto por Frey e Slate (1991) uma determinada regra é descartada caso tenha uma pontuação inferior a um limite estabelecido, e outra regra é gerada para substituí-la. As regras de maior pontuação permanecem no sistema, até que se tenha um conjunto de regras bem definido.

Comparando com outros trabalhos que utilizam o mesmo conjunto de dados, os resultados obtidos por este trabalho foram desfavoráveis ao reconhecimento de caracteres. Porém, este trabalho deve ser considerado importante por ser um dos primeiros voltados para o reconhecimento de padrões caracteres.

3.3.1.2 Schwenk e Bengio (2003)

O trabalho proposto por Schwenk e Bengio (2003) apresenta a aplicação de um algoritmo de reforço adaptativo proposto em Freund e Schphire (1996) aplicado em Redes Neurais Artificiais (Kovács, 2002) voltadas para o reconhecimento de caracteres. Este algoritmo de reforço adaptativo (denominado *AdaBoost*) tem como principal objetivo construir um classificador composto combinando as hipóteses geradas por vários classificadores. Estes classificadores iniciais são treinados dando mais destaque em certos padrões analisados a cada iteração, utilizando uma função de custo ponderado para uma distribuição de probabilidade sobre a base de treinamento e os resultados obtidos pelos classificadores são analisados e tratados matematicamente de forma a gerar um resultado final para o classificador composto. No caso deste trabalho, como todos os classificadores eram baseados em redes neurais artificiais (RNA), foi possível a utilização da mesma função de avaliação para a obtenção destes resultados.

Ao final da aplicação, foi obtido um classificador composto por um conjunto contendo 20 RNAs treinadas para a classificação de caracteres. Para o treinamento destes classificadores e para a obtenção de resultados foram utilizadas duas base de dados, uma contendo padrões de caracteres do alfabeto inglês (a mesma base de dados utilizada no trabalho apresentado nesta monografia) e uma segunda base de dados contendo padrões de dígitos manuscritos (0-9). A principal vantagem observada neste trabalho é a taxa elevada de acertos em ambos os casos, comparando com outros trabalhos que utilizam a mesma técnica aplicada em outras estruturas de classificadores. Porem é importante destacar como principal desvantagem deste classificador o tempo gasto com treinamento das RNA.

3.3.1.3 Husnain e Naweed (2009)

Em Husnain e Naweed (2009) foi utilizado um classificador de caracteres baseado na Teoria de Decisão Bayesiana (*BDT - Bayesian Decision Theory*), que é uma abordagem estatística fundamental para os problemas de classificação genérica de padrões. Segundo esta teoria, a solução para o problema de classificação de padrões é puramente baseado em valores probabilísticos e todos os valores de probabilidade relevantes são conhecidos. Para se obter uma baixa taxa de erros em um classificador, deve-se escolher a hipótese com o menor risco quantificado. Esta teoria é explicada mais detalhadamente no trabalho de Braga (2005), que diz que "a teoria de decisão bayesiana é uma ferramenta que auxilia o processo de escolha, faz uso de probabilidades para a escolha de hipóteses e quantifica o risco existente em tomar uma decisão".

O sistema apresentado possui uma alta eficácia, em comparação com outros trabalhos que utilizam a mesma base de dados. Porém os resultados foram obtidos considerando um conjunto bastante reduzido de padrões de testes, comparado com os demais trabalhos que utilizam o mesmo conjunto de dados.

3.4 Considerações finais

Os sistemas de classificação de padrões são sistemas que separam e classificam objetos ou eventos segundo suas características. Podem ser aplicados em várias áreas, como informática, medicina e educação. Uma possível aplicação para este sistema é a classificação de caracteres em imagens, que pode ser utilizada para digitalizar documentos existentes no meio físico a fim de salvá-los em formatos eletrônicos editáveis. Existem muitos trabalhos na literatura voltados para a classificação de caracteres de vários tipos, como caracteres chineses, árabes, dígitos e caracteres do alfabeto inglês. Foram apresentados três trabalhos que servirão para comparação de resultados com o trabalho proposto por esta monografia que utilizam o mesmo conjunto de dados.

4
Base de dados

Este capítulo apresenta os principais detalhes referentes à base de dados que foi utilizada no desenvolvimento dos modelos de classificação apresentadas neste trabalho.

4.1 Considerações iniciais

Para a realização deste trabalho fez-se necessária a utilização de um conjunto de dados contendo padrões referentes aos caracteres do alfabeto americano (A-Z). Os detalhes, a metodologia de utilização e a análise estatística deste conjunto de dados serão apresentados a seguir.

4.2 Conjunto de dados

Os modelos de classificação de caracteres que são apresentados neste trabalho foram desenvolvidos considerando uma base de dados obtida a partir do *Machine Learning Repository* (WEB1, 2009). Esta base de dados foi desenvolvida por Dr. Allen V. Hershey, um físico matemático do *U.S. Naval Weapons Laboratory*. Foi apresentada pela primeira vez no trabalho de Frey e Slate (1991).

Este conjunto de dados contém 20000 padrões baseados em vinte fontes do alfabeto romano. Algumas distorções foram aplicadas na geração destes padrões, considerando cinco espessuras diferentes de fonte (simples, dupla, tripla, complexa

e gótica), além de seis estilos diferentes de fonte (itálico, *block*, *script*, Inglês, Italiano e Alemão) (Frey e Slate, 1991).

A geração da base de dados foi realizada da seguinte forma: Primeiramente são geradas, aleatoriamente, 20000 imagens de caracteres seguindo as características de fontes e distorções citadas anteriormente. Feito isto, cada imagem é reduzida para um tamanho de 16x16 pixels. Em seguida é realizada, para cada imagem, a extração de características que irão compor um padrão da base de conhecimento. Cada padrão gerado é composto por 16 atributos numéricos que possui um valor inteiro entre 0 e 15. Este conjunto de atributos representa as características visuais da imagem que são utilizadas para se realizar a classificação de um caractere (*ibidem*). Estes atributos, assim como seu significado, são destacados na Tabela 1. A Figura 2 apresenta exemplos de imagens geradas por este sistema.

Tabela 1: Atributos dos padrões de caracteres da base de dados (Frey e Slate,1991)

Atributo	*Significado*	*Comentário*
1	Posicionamento Horizontal	-
2	Posicionamento Vertical	-
3	Largura da figura (em pixels)	-
4	Altura da figura (em pixels)	-
5	Número de pixels ativos na imagem	-
6	Posição média horizontal de todos os pixels ativos em relação ao centro da imagem dividido pela largura da imagem	O valor deste atributo tende a 0 se o caractere for balanceado à esquerda (Ex: Letra "L")
7	Posição média vertical de todos os pixels ativos em relação ao centro da imagem dividido pela altura da imagem	O valor deste atributo tende a 0 se o caractere for balanceado inferiormente. (Ex. Letra "U")
8	Média do valor do quadrado das distâncias horizontais de pixels	Este atributo terá um valor maior em imagens cujos pixels estão mais separados no sentido horizontal (Ex: Letras "M" e "W")

9	Média do valor do quadrado das distâncias verticais de pixels	Este atributo terá um valor maior em imagens cujos pixels estão mais separados no sentido vertical (Ex: Letras "S" e "Z")
10	O produto médio das distâncias horizontais e verticais para cada pixel ativo;	-
11	O valor médio do quadrado da distância horizontal multiplicado pela distância vertical de cada pixel ativo	Este atributo representa a medida da correlação entre a variação horizontal com a posição vertical;
12	O valor médio do quadrado da distância vertical multiplicado pela distância horizontal de cada pixel ativo	Esta mede a correlação entre a variação vertical com a posição horizontal;
13	O número médio de arestas encontradas fazendo varreduras sistemáticas da esquerda para a direita por todas as posições verticais da imagem;	Definição de aresta: um pixel inativo seguido imediatamente de um pixel ativo
14	A soma das posições verticais de arestas encontradas em 13	-
15	O número médio de arestas encontradas ao fazer varreduras sistemáticas de baixo para cima, sobre todas as posições horizontais da imagem	-
16	A soma das posições horizontais de arestas encontradas em 15	-

Figura 2: Exemplo de imagens dos padrões de caracteres existentes na base de dados (Frey e Slate, 1991)

É possível perceber pela Figura 2 que alguns caracteres são ilegíveis até mesmo para uma pessoa. Isto pode dificultar de forma considerável a classificação de caracteres por um modelo inteligente.

É importante destacar que a semelhança entre padrões existentes nesta base de dados influencia de forma significativa nos resultados obtidos por um sistema de classificação. Por exemplo: Uma letra "A" pode ser representada de várias formas, mas sempre mantendo um comportamento padrão entre cada representação, como é ilustrado na Figura 3. Entretanto, existem padrões que possuem comportamento bastante semelhante e representam caracteres distintos, como é apresentado nas Figuras 5 e 6 pela semelhança entre os padrões dos caracteres "C"e "G" e entre os padrões dos caracteres "O" e "Q", respectivamente. Este fator pode dificultar a classificação do caractere por um modelo.

Vale ressaltar que a representação destes padrões em gráfico de linha se dá unicamente para a verificação da semelhança entre os padrões, não levando em consideração outros fatores. Mesmo que se troquem a ordem dos atributos (conforme mostrado na Figura 4), pode-se perceber que a semelhança entre os padrões e a distância euclidiana entre eles permanecem as mesmas.

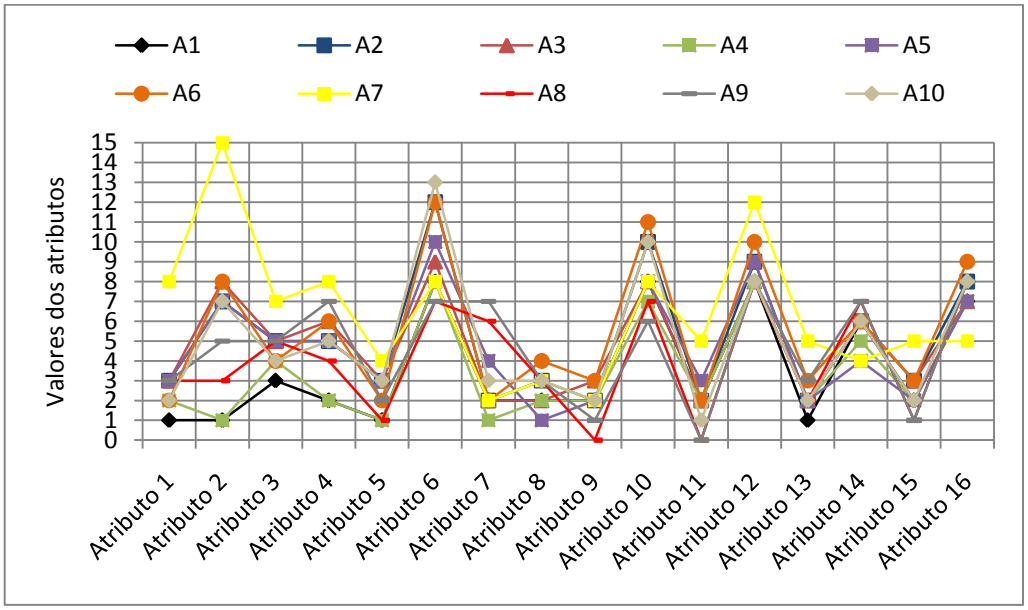

Figura 3: Gráfico referente a dez padrões da letra "A" que compõem a base de dados utilizada neste trabalho.

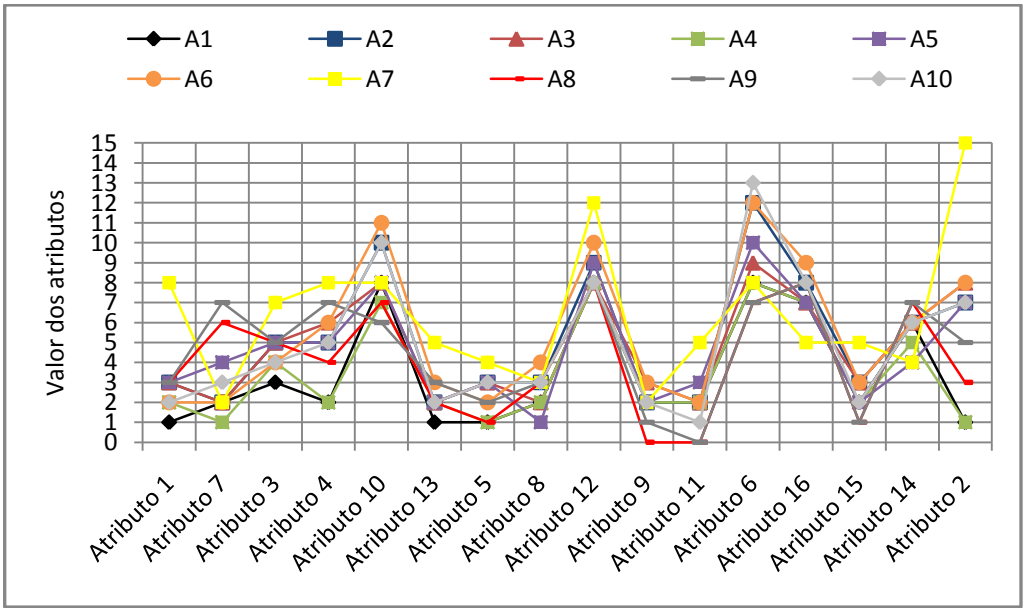

Figura 4: Gráfico referente a dez padrões da letra "A" com os atributos fora de ordem.

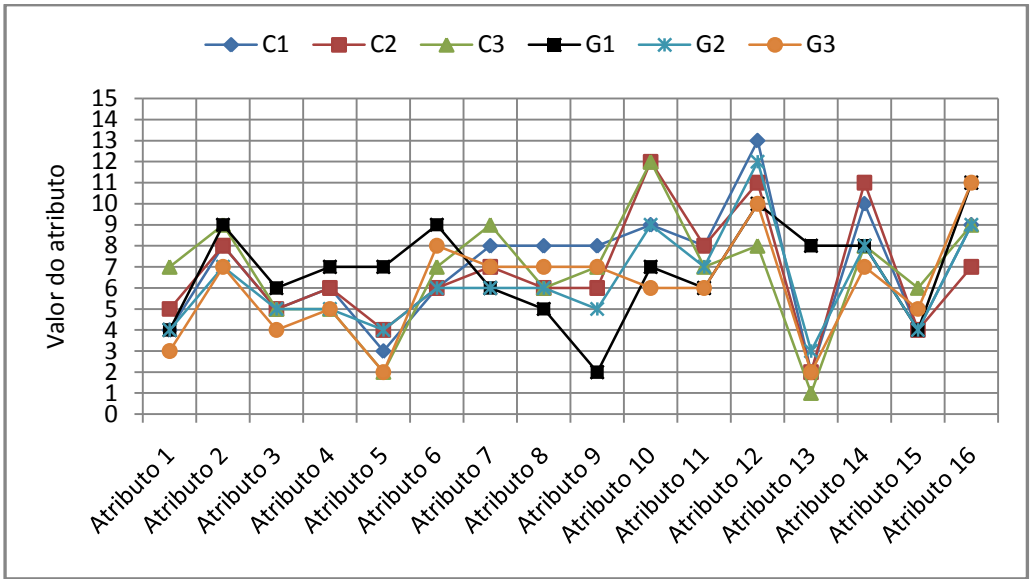

Figura 5: Semelhança entre padrões dos caracteres "C" e "G"

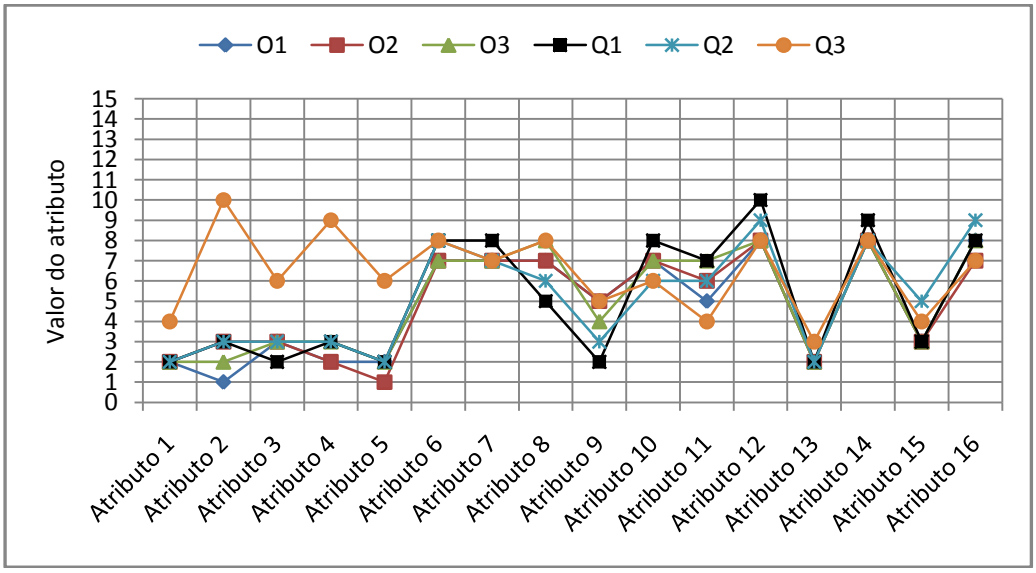

Figura 6: Semelhança entre padrões dos caracteres "O" e "Q"

Para a realização de experimentos, o conjunto de dados foi dividido em duas partes. A primeira é composta pelos primeiros 16000 padrões de caracteres e é utilizada como base de conhecimento (BC) dos sistemas propostos. A segunda parte é utilizada na realização dos testes e obtenção de resultados e é denominada base de testes (BT) e é composta pelos 4000 padrões restantes. A metodologia de separação da base de dados foi a mesma utilizada nos trabalhos de Frey e Slate

(1991), Schwenk e Bengio (2003) e Husnain e Naweed (2009), os quais foram utilizados para a comparação de resultados com este trabalho.

A Tabela 2 apresenta a distribuição dos padrões na base de dados, contemplando, para cada caractere, o total contido na base de dados (T), a porcentagem em relação ao conjunto de dados completo (%T1), o total contido na base de conhecimento (BC), sua porcentagem em relação ao conjunto de dados completo (%T2) e em relação à base de conhecimento (%BC), o total contido na base de testes (BT), sua porcentagem em relação ao conjunto de dados completo (%T3) e em relação à base de testes (%BT).

Tabela 2: Distribuição dos padrões de caracteres existentes na base de dados

Caractere	*T*	*%T1*	*BC*	*%T2*	*BC (%)*	*BT*	*%T3*	*BT (%)*
A	789	3,945	633	3,165	3,956	156	0,780	3,900
B	766	3,830	630	3,150	3,938	136	0,680	3,400
C	736	3,680	594	2,970	3,712	142	0,710	3,550
D	805	4,025	638	3,190	3,988	167	0,835	4,175
E	768	3,840	616	3,080	3,850	152	0,760	3,800
F	775	3,875	622	3,110	3,887	153	0,765	3,825
G	773	3,865	609	3,045	3,806	164	0,820	4,100
H	734	3,670	583	2,915	3,644	151	0,755	3,775
I	755	3,775	590	2,950	3,687	165	0,825	4,125
J	747	3,735	599	2,995	3,744	148	0,740	3,700
K	739	3,695	593	2,965	3,706	146	0,730	3,650
L	761	3,805	604	3,020	3,775	157	0,785	3,925
M	792	3,960	648	3,240	4,050	144	0,720	3,600
N	783	3,915	617	3,085	3,856	166	0,830	4,150
O	753	3,765	614	3,070	3,838	139	0,695	3,475
P	803	4,015	635	3,175	3,969	168	0,840	4,200
Q	783	3,915	615	3,075	3,844	168	0,840	4,200
R	758	3,790	597	2,985	3,731	161	0,805	4,025
S	748	3,740	587	2,935	3,669	161	0,805	4,025
T	796	3,980	645	3,225	4,031	151	0,755	3,775
U	813	4,065	645	3,225	4,031	168	0,840	4,200
V	764	3,820	628	3,140	3,925	136	0,680	3,400
W	752	3,760	613	3,065	3,831	139	0,695	3,475
X	787	3,935	628	3,140	3,925	159	0,795	3,975
Y	786	3,930	641	3,205	4,006	145	0,725	3,625
Z	734	3,670	576	2,880	3,600	158	0,790	3,950

A fim de se analisar a relação entre a configuração da base de dados e os resultados a serem obtidos foram geradas outras 59 configurações diferentes de base de conhecimento e testes, mantendo a quantidade de padrões dos conjuntos originais, sendo que 55 destas configurações foram geradas de forma puramente aleatória e as 4 restantes foram geradas de forma pseudo-aleatórias (PA), mantendo-se a mesma proporção de caracteres existente na configuração original.

Tabela 3: Análise estatística dos dados

Base de dados	*Média*	*Variância*	*Coeficiente de variação (%)*	*Desvio padrão*
Base de dados completa	5,925466	8,455712	49,07413	2,907871
BC Original	5,924445	8,472864	49,13234	2,910818
BT Original	5,924547	8,387214	48,84132	2,896069
BC PA - 1	5,925664	8,45821	49,07974	2,9083
BT PA - 1	5,924672	8,445848	49,05207	2,906174
BC PA - 2	5,925816	8,447206	49,04654	2,906408
BT PA - 2	5,924063	8,489866	49,18417	2,913737
BC PA - 3	5,927395	8,447718	49,03497	2,906496
BT PA - 3	5,91775	8,487743	49,23109	2,913373
BC PA - 4	5,926641	8,450433	49,04908	2,906963
BT PA - 4	5,920766	8,476932	49,17467	2,911517
BC Aleatória	5,927434	8,422568	48,9616	2,902166
BT Aleatória	5,917594	8,588343	49,5233	2,930588

Com base nas medidas de posição e variabilidade apresentadas na Tabela 3, acredita-se que as variabilidades das BCs e BTs são equivalentes. Neste sentido, podem-se utilizar as configurações apresentadas anteriormente para medir o desempenho dos modelos propostos neste trabalho.

4.3 Considerações finais

Neste capítulo foi apresentada a base de dados utilizada neste trabalho para a realização de experimentos. Este conjunto de dados possui 20000 padrões de caracteres do alfabeto inglês (A-Z) baseados em 20 tipos diferentes de fontes do alfabeto romano e seguindo um conjunto de formas e distorções. Este conjunto foi separado em duas partes, a primeira contendo os primeiros 16000 padrões, considerada base de conhecimento, e a segunda contendo os restantes 4000 padrões, denominada base de testes.

Além da configuração original, foram geradas 59 configurações diferentes de bases de conhecimento e testes à fim de avaliar a relação entre a configuração da base de dados e os resultados obtidos, entre elas 4 configurações que possuem a mesma proporcionalidade entre os caracteres que a configuração original e as demais configurações geradas de forma puramente aleatória. Todas as configurações geradas sempre mantêm a definição de 16000 padrões para a BC e 4000 para a BT.

5

Proposta

Este capítulo apresenta a proposta desta monografia. É organizado da seguinte forma: A Seção 5.1 apresenta uma breve introdução do capítulo. A Seção 5.2 apresenta uma breve descrição sobre Sistemas Baseados em conhecimento. Na Seção 5.3 serão apresentados os modelos propostos. Encerrando o capítulo, na Seção 4.4 são apresentadas as considerações finais.

5.1 Considerações iniciais

Com base no conjunto de dados descrito no capítulo 4, este trabalho propõe modelos de classificação capazes de classificar caracteres reconhecidos através de uma imagem digitalizada de um documento. Serão apresentados neste capítulo os dois modelos propostos, sendo dois destes baseados em Sistemas Especialistas (SE) e o terceiro modelo baseado na técnica de Máquinas de Vetor de Suporte. A criação de três modelos distintos deve-se apenas para comparação de resultados obtidos entre eles e outras propostas da literatura.

5.2 Sistemas baseados em conhecimento

Ambos os sistemas propostos por este trabalho são sistemas que utilizam o conhecimento representado de forma explícita para resolver problemas, sendo

assim denominados Sistemas Baseados em Conhecimento (SBC). Este tipo de sistema manipula o conhecimento e as informações de forma inteligente e geralmente são desenvolvidos para resolver problemas que necessitam de uma quantidade considerável de especialização e conhecimento humano. Assim, conhecimento e processo de resolução de problemas são os principais pontos para se desenvolver um sistema baseado em conhecimento. (Rezende, 2005).

Os sistemas baseados em conhecimento, segundo Jackson (1998) *apud* Rezende (2005, p16), devem ser capazes de:

- "Questionar o usuário, utilizando uma linguagem de fácil entendimento, para reunir as informações necessárias";
- "Desenvolver uma linha de raciocínio a partir dessas informações e do conhecimento nele embutido para encontrar soluções satisfatórias";
- "Explicar seu raciocínio, caso seja questionado pelo usuário, o porquê necessita de informações externas e de como chegou às suas conclusões";
- "Conviver com seus erros, i.e., tal como um especialista humano, os SBCs podem cometer erros, mas devem possuir um desempenho satisfatório que compense seus possíveis equívocos".

Segundo Motta (1998) *apud* Rezende (2005), Tudo o que um SBC "souber" sobre um determinado problema deve estar armazenado de forma explícita na base de conhecimento, que deverá ser utilizada por um agente que saiba interpretá-la. Geralmente os SBC resolvem problemas onde não é conhecido um procedimento determinístico que garanta uma resolução efetiva. Estes sistemas utilizam conhecimento específico do domínio visando contornar a exponencialidade de formulação genérica do problema e/ou a ausência de conhecimento preciso e completo sobre seu domínio.

Os SBC propostos neste trabalho trabalham sobre as 16 características que compõem um padrão de caractere, analisando-os considerando a BC do conjunto

de dados e, feito isto, apresentando a classificação do padrão tratado. A Figura 7 apresenta um exemplo de funcionamento de um SBC.

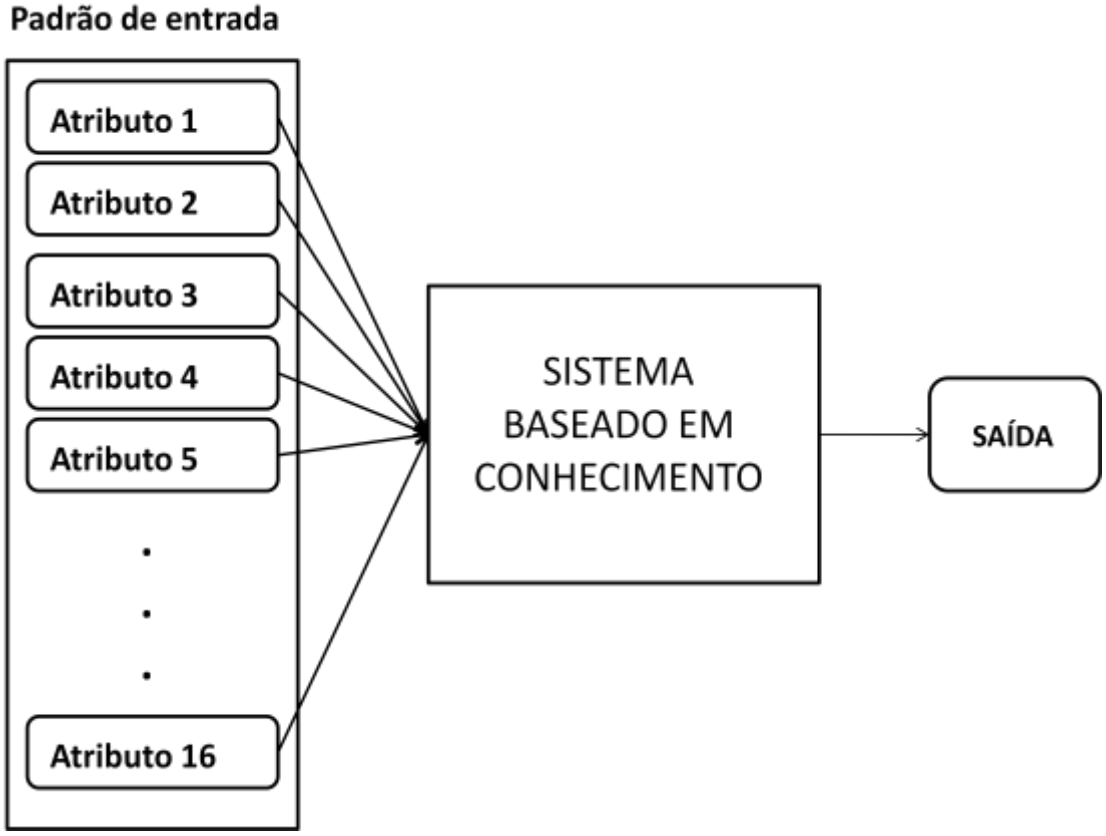

Figura 7: Funcionamento de um SBC

Muitos trabalhos na literatura apresentam uma nomenclatura errônea sobre os SBC, definindo-os como sistemas especialistas. É importante fazer a diferenciação destes dois termos. Os SBC são capazes de resolver problemas utilizando conhecimento específico da aplicação, enquanto os sistemas especialistas são SBC que resolvem problemas que podem ser resolvidos por conhecimento humano, requerendo assim informações sobre a habilidade, a experiência e as técnicas utilizadas pelo especialista.

A seguir serão apresentados os SBC propostos neste trabalho para a classificação de caracteres. O primeiro modelo se baseia em Sistemas Especialistas.

O segundo modelo trata-se de uma variação do primeiro e o terceiro modelo é baseado na técnica de Máquinas de Vetores de Suporte (SVM).

5.2.1 Sistema Especialista

Dentre as várias abordagens estudadas na área de inteligência artificial para problemas de classificação de padrões podem ser destacados os sistemas chamados sistemas especialistas ou sistemas baseados em regras. Os sistemas especialistas (SE) são sistemas que analisam informações sobre um determinado evento ou classe específica e trabalha com estas informações tendo como suporte um conjunto de regras. A utilização de um sistema especialista (SE) está voltada para uma determinada aplicação do conhecimento humano, sendo este limitado. Geralmente são problemas que podem ser solucionados por um especialista humano. Estes especialistas fornecem regras gerais indicando como analisariam o problema e uma nova informação de entrada seria analisada de acordo com as regras definidas.

Um SE simples geralmente possui uma arquitetura com três componentes: um conjunto de regras na forma "*se* condição *então* ação", uma memória de trabalho e um motor de inferência. O conjunto de regras e a memória de trabalho formam a base de conhecimento do SE, onde é representado o conhecimento do sistema sobre o domínio. O motor de inferência é o componente de controle do sistema especialista que irá lidar com decisões baseando-se no conjunto de regras.

Com a entrada de novos dados, o sistema então passa para um processo de raciocínio no qual vai formulando novas hipóteses e verificando novos fatos, nos quais vão influenciar no processo de raciocínio. Este raciocínio é sempre baseado no conhecimento prévio acumulado. Caso o fato já previamente conhecido pelo SE não seja suficiente, a margem de erro se torna maior, chegando até mesmo numa conclusão errada. No entanto vale ressaltar que este erro é justificado em função dos fatos que encontrou e o do seu conhecimento acumulado previamente. Logo,

um SE além de obter conclusões, deve ter a capacidade de aprender novos conhecimentos, melhorando o seu desempenho de raciocínio e qualidade de suas decisões.

A Figura 8 mostra o modelo do SE apresentado neste trabalho.

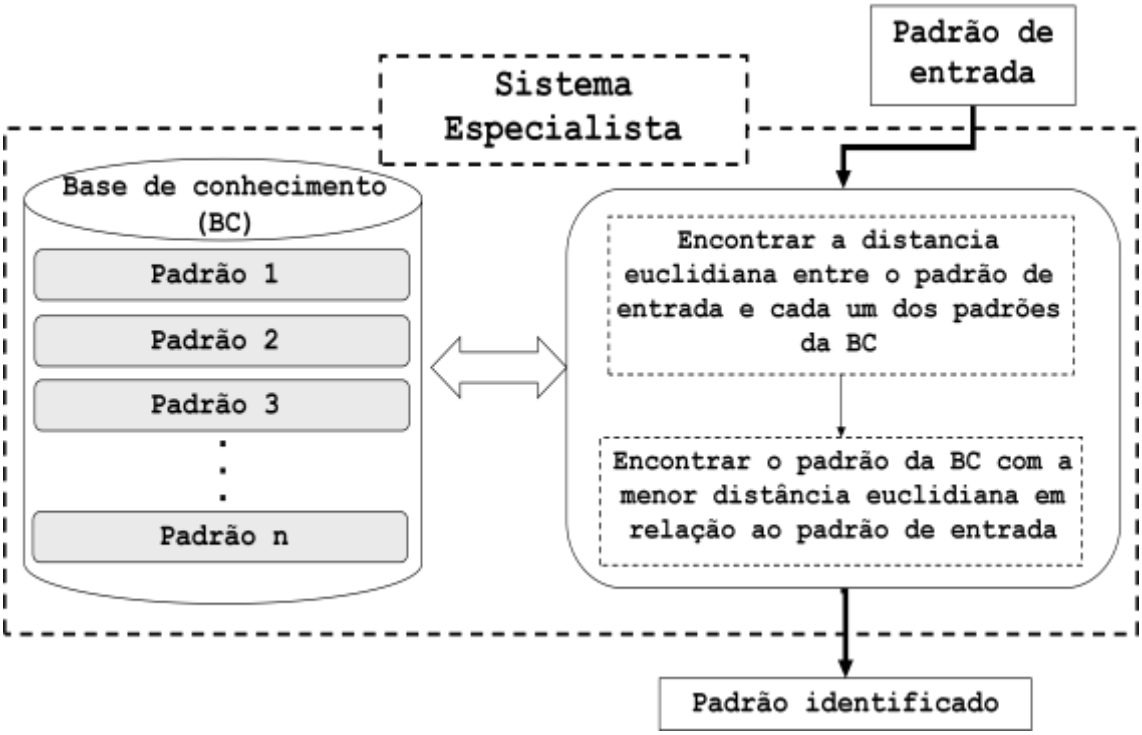

Figura 8: Modelo do sistema especialista apresentado neste trabalho

O SE proposto funciona da seguinte maneira: primeiramente, cada elemento da base de testes é comparado com todos os elementos da base de conhecimento utilizando a fórmula da distância euclidiana descrita por

$$\boldsymbol{D}\,(\boldsymbol{X},\boldsymbol{Y}) = \sqrt[2]{\sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{y_i})^2} \qquad (\mathbf{1})$$

onde X representa a base de conhecimento, Y representa a base de testes, n é número de características de um padrão, x_i equivale à i-ésima característica do padrão da base de conhecimento e y_i equivale à i-ésima característica do padrão da base de testes.

O conjunto de regras deste sistema especialista possui somente uma regra: *SE* a distância euclidiana entre o padrão de entrada e o padrão da base de conhecimento for mínima (ou seja, se for a menor distância euclidiana encontrada para todos os padrões da base de conhecimento), *ENTÃO* estes padrões pertencem à mesma classe. Em outras palavras, o sistema especialista desenvolvido considera que, quanto menor a distância euclidiana entre dois padrões quaisquer, maior a semelhança entre suas imagens, o que implica na maior a probabilidade destes padrões pertencerem à mesma classe.

Após a realização deste cálculo entre cada padrão da base de testes mediante a base de conhecimento, obtém-se uma matriz de valores de distância. Estas distâncias são ordenadas de forma crescente e o índice do padrão da base de conhecimento que possuir a menor distância euclidiana em relação ao padrão da base de teste indicará qual é a letra reconhecida, classificando-a de acordo com os padrões da base de conhecimento.

5.2.1.1 Modificação

Visando reduzir o problema de similaridade entre padrões citados no capítulo 4, foi desenvolvida uma variação deste SE onde se obtêm os K padrões que possuem as menores distâncias euclidianas em relação ao padrão de entrada (o valor de K é definido pelo usuário do sistema) e a classificação é realizada com base no índice de maior ocorrência dentre os padrões obtidos. Lembrando que este método só é executado se a classificação de um determinado padrão for equivocada, considerando somente a menor distância euclidiana encontrada. A Figura 9 apresenta o funcionamento desta modificação.

Apesar de ser um método muito simples, o modelo de SE proposto, assim como sua variação, se comportou de forma bastante satisfatória no que diz respeito à taxa de acerto e, principalmente, à velocidade de execução. Esta análise será apresentada no capítulo 5.

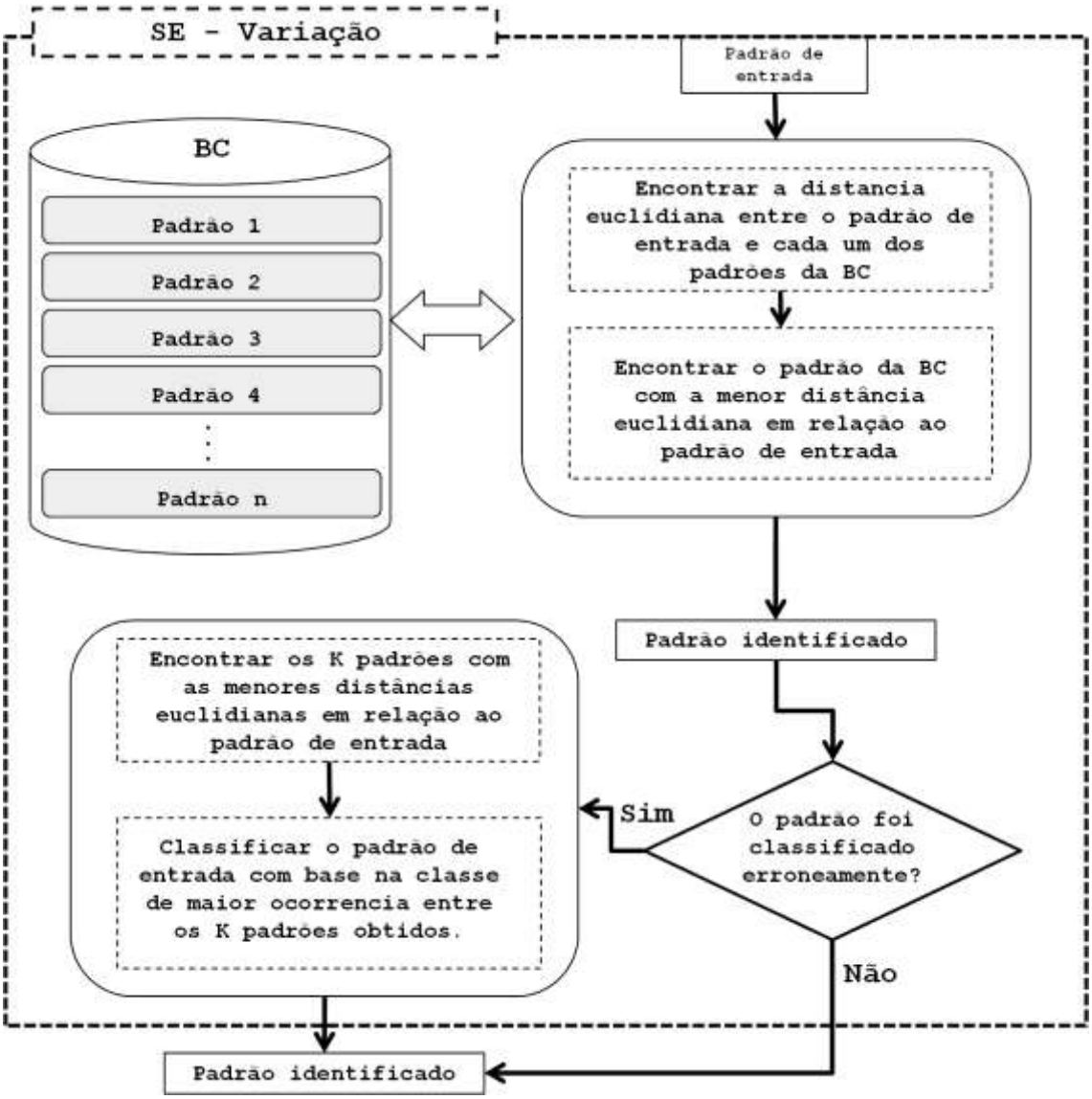

Figura 9: Modificação no SE proposto

5.2.2 Maquina de Vetores de Suporte (SVM)

A Máquina de Vetores de Suporte *(Support Vector Machines* – SVM) é uma técnica baseada na teoria de aprendizagem estatística sugerida inicialmente por Vapnik (1998). É uma ferramenta de classificação amplamente utilizada no meio científico e que, segundo Semolini (2002), "em poucos anos desde que foi introduzida, já

apresenta um desempenho superior à maioria dos outros métodos em uma ampla variedade de aplicações".

O objetivo principal do SVM é realizar um mapeamento não-linear do conjunto de dados de entrada para um espaço característico de alta-dimensão, onde um hiperplano de separação é obtido para que se possam resolver os problemas de classificação. De acordo com Baranoski, Justino e Bortolozzi (2005),

> A técnica se baseia no princípio da Minimização do Risco Estrutural (MRS). O princípio da indução do MRS possui dois objetivos. O primeiro é controlar o risco empírico no conjunto de treinamento. O segundo é controlar a capacidade da função de decisão usada para obter esse valor de risco.

O MRS é profundamente baseado na dimensão de Vapnik-Chervonenkis (VC), que é uma medida da capacidade de classificação de um grupo de funções indicadoras calculadas previamente por uma máquina de aprendizagem, onde o seu valor é equivalente ao número máximo de exemplos de treinamento que podem ser aprendidos sem erros. O valor da dimensão VC é equivalente ao número máximo de exemplos de treinamento que podem ser aprendidos sem erros. Pode ser representado pela equação $VC = n + 1$, sendo n a dimensão do espaço vetorial em questão, ou seja, o número de vetores que podem ser separados em duas classes diferentes, para uma dimensão especificada. Por exemplo, VC=2 quando o problema pode ser separado por, no mínimo, um hiperplano, VC=3 quando o problema pode ser separado por, no mínimo, dois hiperplanos e assim por diante. A Figura 10 apresenta um exemplo de dimensão VC.

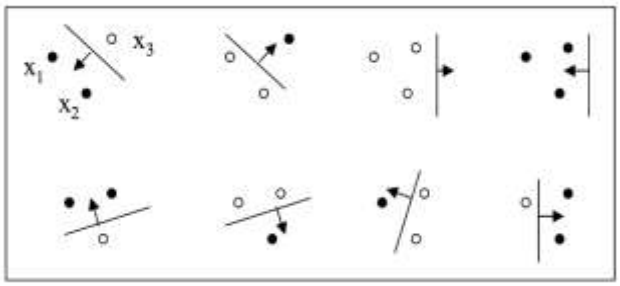

Figura 10: Dimensão VC. Neste caso: VC = 3 (Semolini, 2002)

É importante destacar que a técnica SVM é dividida em duas vertentes, pois embora tenha sido criada para atender casos em que as saídas são números discretos (classificação), ela foi modificada para também atender casos em que as saídas são números contínuos (previsão). Neste trabalho é utilizada a técnica SVM para classificação, denominada SVC (*Support Vector Classifier*). No princípio, se utilizava o SVC para classificação binária (entre duas classes). Atualmente o SVC pode ser aplicado para a classificação de padrões em n-classes. Segundo Semolini (2002), a flexibilidade é um ponto forte do SVM, pois "pode-se adaptar o problema de classificação binária (apenas com duas classes), que foi a abordagem que originou a formulação da SVM, para resolver muitos outros tipos de problemas."

Um conceito fundamental no projeto de SVMs é o conceito de margem de separação entre classes associadas à taxa de erro permitida na classificação. Esta margem representa a menor distância entre os exemplos do conjunto de treinamento e o hiperplano utilizado para a separação destes conjuntos em classes (Lorena e Carvalho, 2007). O SVM trabalha focando na maximização da margem, i.e., Quanto maior o valor da margem, melhor duas classes podem ser separadas.

Outro conceito fundamental é o conceito de vetores suporte. Esses vetores são padrões críticos que sozinhos determinam o hiperplano ótimo (i.e., o hiperplano que possui o maior valor de margem possível), enquanto os demais padrões (não-críticos) são considerados irrelevantes e são desconsiderados para a solução do problema de separação. Informalmente, os vetores suporte são os padrões que possuem a menor distância do valor da margem. Na Figura 11 é apresentado exemplos de hiperplano ótimo, margem e vetores de suporte, onde as linhas coloridas (Figura 11a) representam os hiperplanos encontrados pelo SVM, W_1 e W_2 representam os conjuntos diferentes de padrões e os padrões preenchidos com cores mais claras (Figura 11b) representam os vetores de suporte, que se encontram a uma distância mínima δ do hiperplano ótimo.

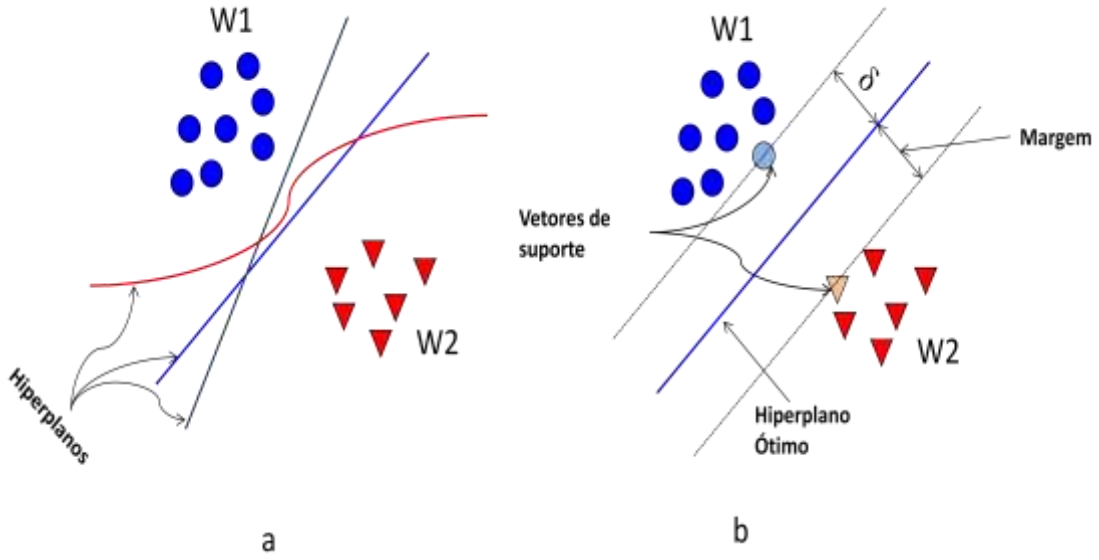

Figura 11: Exemplo de hiperplanos, hiperplano ótimo, margem e vetores de suporte em SVM

Existem situações onde o hiperplano encontrado não realiza corretamente a divisão de duas classes utilizando uma função linear, conforme mostrado na Figura 11. Para isto, deve-se realizar uma modificação no *kernel,* que é a função que representa o hiperplano encontrado durante o processo de separação das classes. Além da função linear, outras funções são comumente utilizadas em projetos SVM. A Tabela 4 apresenta alguns exemplos de funções *kernel*.

Tabela 4: Exemplos de kernel para SVM. (WEB1, 2009);

Função	***Equação***
Linear	$k(x_i, x_j) = x_i \cdot x_j$
Polinomial	$k(x_i, x_j) = (x_i * x_j + c)^d$
Base Radial	$k(x_i, x_j) = \exp\left(-\gamma \left\|\|x_i - x_j\right\|\|^2\right), para\ \gamma > 0$
Sigmoidal	$k(x_i, x_j) = tahn(\gamma \cdot x_i \cdot x_j + c), para\ \gamma > 0\ e\ c < 0$

De acordo com Rezende (2005), o problema de separação de classes depende dos dados de treinamento, dos parâmetros que definem a *kernel* e da margem obtida na separação, sendo formulado como um problema de programação quadrática com restrições lineares. No caso do SVM, trata-se de um problema

determinístico, ou seja, a mesma base de dados, os mesmos dados de treinamento e o mesmo *kernel* devem resultar na mesma resposta.

Assim como o sistema especialista apresentado anteriormente, o SVM trabalha sobre os atributos dos padrões da BC para a definição dos hiperplanos, vetores de suporte e margens. Em seguida, cada padrão de entrada é classificado de acordo com os conjuntos definidos pelos hiperplanos gerados. A Figura 12 apresenta o funcionamento do modelo SVM utilizado neste trabalho.

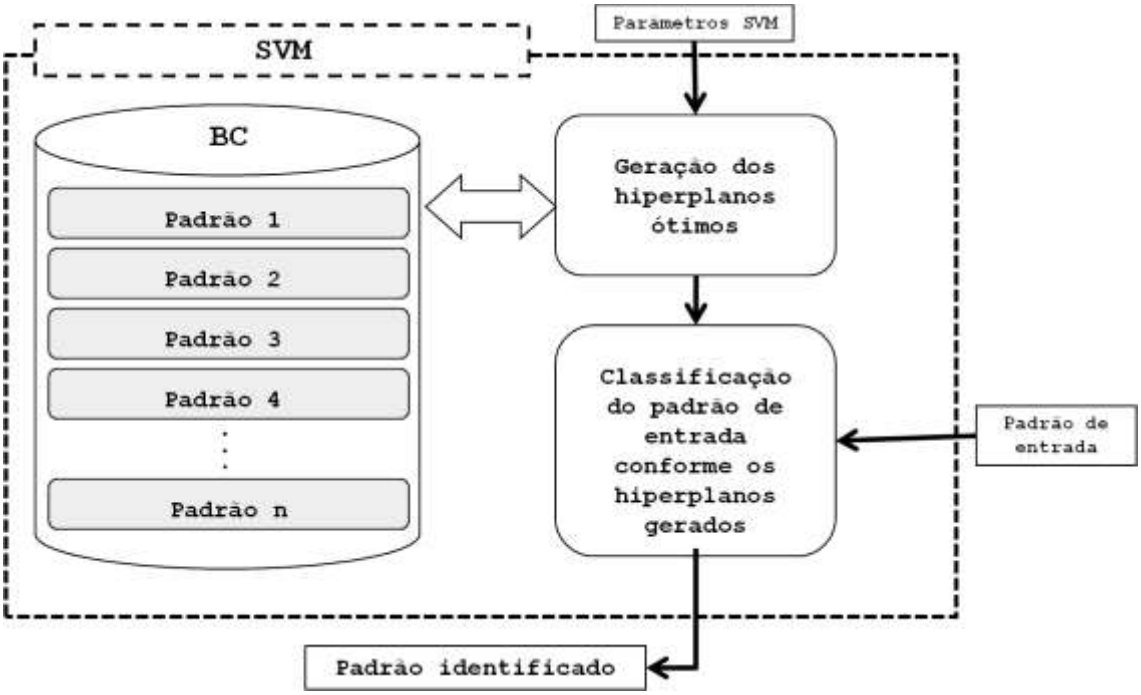

Figura 12: Modelo do sistema SVM utilizado neste trabalho

A utilização do SVM neste trabalho para a classificação de caracteres atendeu as expectativas, no que diz respeito à taxa de acerto na classificação. Detalhes sobre esta análise serão apresentados no Capítulo 5.

5.3 Considerações finais

Este trabalho de conclusão de curso propõe modelos inteligentes aplicados na resolução do problema de classificação de caracteres. Foram apresentadas duas abordagens distintas que foram aplicadas para a resolução do problema de classificação de caracteres proposto. A primeira delas utiliza um sistema especialista que é baseado no calculo da distância euclidiana entre os padrões da base de conhecimento e de teste. A segunda abordagem é baseada na técnica de SVM, que tem como objetivo gerar hiperplanos que realize a separação de classes distintas.

6
Resultados

Este capítulo apresenta os resultados obtidos pelos sistemas de classificação de caracteres apresentados neste trabalho. Além da análise dos resultados, também será apresentada uma comparação com outras propostas da literatura, descritas no capítulo 3.

6.1 Considerações iniciais

Neste capítulo serão apresentados, de forma separada, os experimentos realizados e os resultados obtidos pelo sistema especialista proposto e pela utilização do SVM, respectivamente. Em seguida será realizada a comparação dos resultados obtidos por este trabalho com os trabalhos já citados anteriormente. Ao final do capítulo será mostrado um comparativo com outros trabalhos da literatura.

6.2 Resultados – Sistema Especialista

A obtenção de resultados pelo SE proposto ocorreu da seguinte maneira: Primeiramente se obtém os resultados referentes à execução o sistema para todas as configurações da base de dados, considerando somente a menor distância euclidiana entre padrões da base de conhecimento e da base de teste. Feito isso, se incrementa o número de distâncias mínimas a serem consideradas e se executa o sistema novamente. Informalmente, na segunda execução do sistema são consideradas as duas menores distâncias euclidianas entre os padrões; na terceira execução se consideram as três menores distâncias, e assim por diante, até que o

sistema seja executado 160 vezes, para todas as configurações de dados. Este valor máximo foi escolhido por representar 1% da base de conhecimento definida.

A Tabela 5 apresenta os resultados obtidos com a classificação de caracteres realizadas pelo sistema especialista apresentado neste trabalho utilizando-se a configuração original e as quatro configurações pseudo-aleatórias para a base de dados. Foram considerados os resultados da execução com somente uma distância euclidiana (TA-1), a execução com 160 distâncias euclidianas (TA-160), a média e desvio padrão de todas as 160 execuções realizadas e, para cada base, o número de acerto máximo obtido nas 160 execuções (MAX).

Tabela 5: Resultados obtidos pelo SE proposto executado para as configurações original e pseudo-aleatórias da base de dados

Configuração	*TA-1*	*TA-160*	*Média de acertos*	*Desvio padrão*	*MAX*
Original	95,65 %	97,025 %	97,33 %	9,5598	97,65 %
PA-1	95,15 %	96,675 %	96,914 %	10,3921	97,25 %
PA-2	96,05 %	96,975 %	97,358 %	12,3109	97,925 %
PA-3	95,175 %	96,675 %	96,943 %	10,2452	97,37 %
PA-4	95,175 %	96,375 %	96,806 %	11,2899	97,325 %

Pela Tabela 5 pode-se perceber que o SE proposto conseguiu uma elevada taxa média de acertos na classificação de caracteres para todas as configurações mostradas. Os resultados obtidos com a execução do SE para as configurações pseudo-aleatórias indicam que a configuração da base de dados influencia diretamente no resultado da classificação por este sistema. Podemos perceber também que o sistema apresenta uma maior eficácia na classificação de caracteres à medida que aumentamos o número de distâncias euclidianas consideradas para a classificação. Porém, se o número de distâncias consideradas foi demasiadamente elevado, o sistema pode não possuir uma eficácia desejada. Este fato é comprovado na Figura 13, onde é destacada a quantidade de acerto para cada base de dados em relação ao número de distâncias euclidianas consideradas para a classificação dos padrões de caracteres.

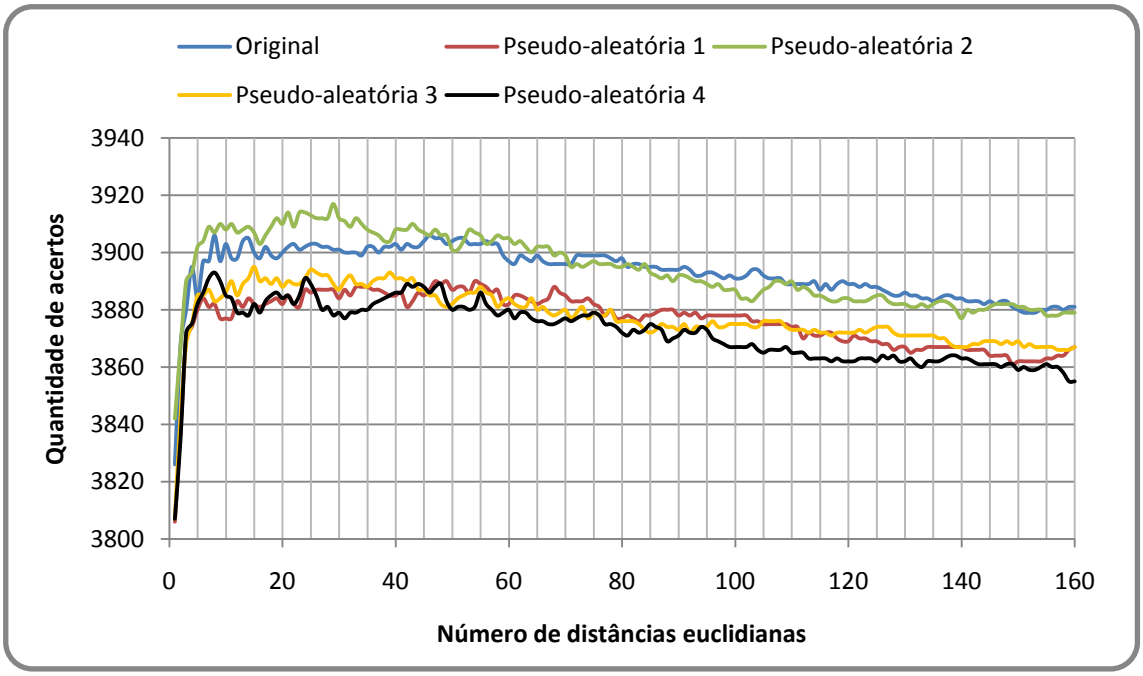

Figura 13: Quantidade de acertos por número de distâncias euclidianas consideradas para a classificação de caracteres.

As demais configurações para a base de dados foram geradas com o intuito de medir um valor de distâncias euclidianas que devem ser consideradas na execução do sistema especialista a fim de maximizar a taxa de acerto. A Tabela 6 apresenta a média de acerto entre todas as bases, a média dos valores máximos atingidos (MMAX), o valor máximo geral atingido (MAX), a média da taxa de acerto executando o sistema considerando somente uma distância euclidiana (M1) e com 160 distâncias euclidianas (M160) para as configurações aleatórias da base de dados.

Tabela 6: Taxas de acertos do sistema considerando execuções utilizando 55 bases aleatórias

Média de acerto geral	*MMAX*	*MAX*	*M1*	*M160*
97,352%	**97,781%**	**98,225%**	**95,705%**	**97,141%**

É possível perceber que o sistema se comportou de forma eficaz, no que diz respeito à classificação de caracteres, para as configurações aleatórias da base de dados. Com a execução deste sistema considerando estas configurações, pode-se

obter uma faixa de valores ideal para se definir quantas distâncias euclidianas devem ser consideradas. A Figura 14 apresenta a relação entre o número de distâncias consideradas para cada configuração aleatória e a quantidade de vezes em que este valor atingiu o melhor resultado obtido dentre as 160 execuções.

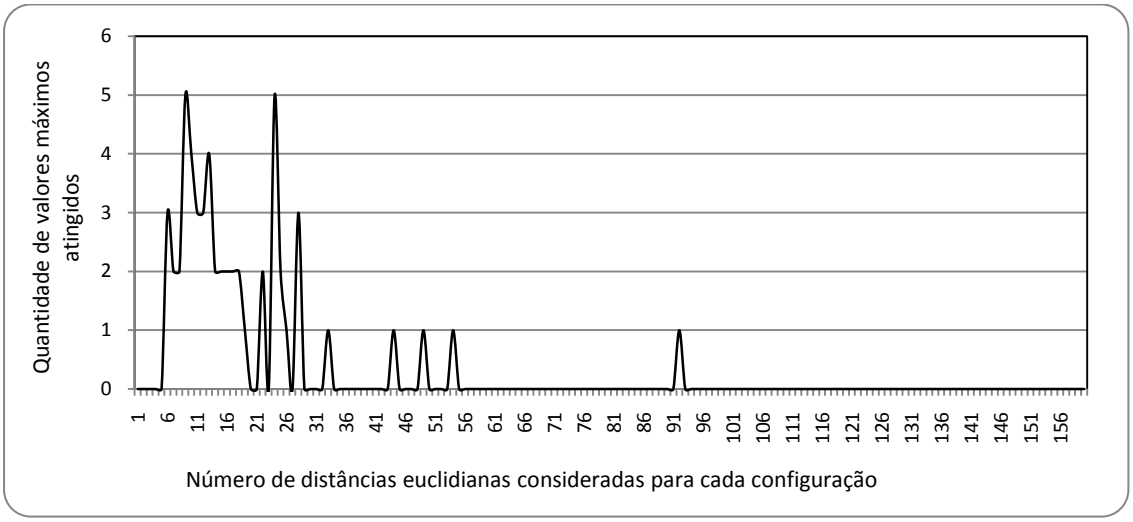

Figura 14: Relação entre o número de distâncias euclidianas consideradas para as execuções em cada configuração da base de dados e o número de acertos máximos obtidos

Pela Figura 14 é possível notar que a faixa onde de distâncias euclidianas onde se encontra as taxas de acertos máximas obtidas está no intervalo entre 6 e 31 distâncias, para as 55 configurações aleatórias da bases de dados utilizada. Por se tratar de um fator aleatório, não é possível afirmar com firmeza se esta relação possa ser considerada para outras bases de dados quaisquer. Este fator depende diretamente da distribuição dos padrões nas configurações da base de dados.

Pode-se concluir que o sistema especialista proposto realiza a classificação de caracteres obtendo uma taxa de acertos satisfatória. Mesmo com a taxa média de acertos elevada, pode-se considerar como principal vantagem encontrada neste sistema a velocidade de execução, pelo fato de não ser necessária a realização de nenhum tipo de treinamento. Como desvantagem pode-se destacar o fato de o classificador considerar dois padrões como sendo da mesma classe somente através da distância euclidiana entre eles, o que nem sempre ocorre como já descrito no capítulo 4. Outra desvantagem encontrada é a falta de uma técnica específica para

definir qual a quantidade ideal de distâncias euclidianas que devem ser consideradas para qualquer configuração da base de dados.

6.3 Resultados – SVM

Para a realização de experimentos utilizando o SVM foi utilizada a ferramenta LibSVM, desenvolvida por Chang e Lin (WEB2, 2010). A escolha da função *kernel*, assim como a configuração de seus parâmetros, foi realizada uma busca exaustiva, onde foram executadas 1000 chamadas da função SVM definindo o *kernel* e seus parâmetros aleatoriamente para cada execução, utilizando somente a configuração original da base de dados. Feito isto, obteve-se as 30 melhores configurações de parâmetros do sistema (i.e., as configurações que possuíram as taxas de acerto mais elevadas) e estas configurações foram armazenadas para serem utilizadas no treinamento e teste das demais configurações da base de dados.

A Tabela 7 apresenta o resultado da classificação de caracteres realizada pela ferramenta SVM. A taxa de acerto considerada equivale à média das taxas de acertos das 30 execuções realizadas para a base de dados original e para as bases de dados pseudo-aleatórias.

Tabela 7: Resultados da execução do SVM para as configurações original e pseudo-aleatórias da base de dados

Base de dados	*Média de acertos*	*Desvio padrão*	*Taxa máxima de acerto alcançada*	*Taxa mínima de acerto alcançada*
Original	95,824%	1,1593	97.9%	95%
PA-1	95,763%	1,0177	97,675%	94,65%
PA-2	95,969%	1,0366	97,85%	95,125%
PA-3	95,625%	1,2217	97,85%	94,75%
PA-4	96,1592%	0,8429	97,7%	95,1%

A Tabela 8 apresenta o resultado da classificação de caracteres realizada utilizando as 55 bases de dados geradas aleatoriamente. É considerada a média

geral de acertos de todas as execuções realizadas, a média das taxas máximas de acerto encontradas para cada configuração da base de dados (MMAX), a média das taxas mínimas de acerto encontradas para cada configuração da base de dados (MMIN) e as taxas máximas (MAX) e mínimas (MIN) de acerto encontradas em todas as execuções do sistema.

Tabela 8: Resultados obtidos pela classificação de caracteres por SVM utilizando as configurações aleatórias da base de dado

Média das taxas de acertos	*MMAX*	*MMIN*	*MAX*	*MIN*
95,8491%	97,8764%	94,9627%	98,325%	94,3%

O sistema SVM também se comportou de forma bastante eficaz no que diz respeito à classificação de caracteres. Assim como no sistema especialista, é possível notar que a configuração da base de dados também influencia, de forma considerável, nos resultados obtidos pelo sistema SVM. Isto é confirmado pelo fato de que a máxima taxa de acerto obtida por ambos os sistemas ocorreu em outras configurações da base de dados, e não na configuração original.

A seguir será apresentada uma comparação feita entre os resultados obtidos com este trabalho e as propostas da literatura descritas anteriormente no capítulo 3.

6.4 Comparativo

A fim de analisar a eficácia do sistema desenvolvido foi realizada uma comparação dos resultados obtidos pelos métodos de classificação propostos com os resultados obtidos dos trabalhos de Frey e Slate (1991), Schwenk e Bengio (2003) e Husnain e Naweed (2009). Para esta comparação foram consideradas somente as execuções realizadas utilizando-se a configuração original, a fim de manter a similaridade entre as configurações da base de dados dos trabalhos citados acima.

A Figura 15 apresenta esta comparação. Para o SE proposto neste trabalho foram considerados os resultados obtidos se considerando somente a menor distância euclidiana, as 160 menores distâncias euclidianas, a média geral de acerto e o valor máximo de acerto obtido em todas as execuções do sistema. Para a utilização do SVM foram consideradas as taxas média, máxima e mínima de acerto entre todas as configurações de parâmetros utilizadas para a realização da obtenção de resultados.

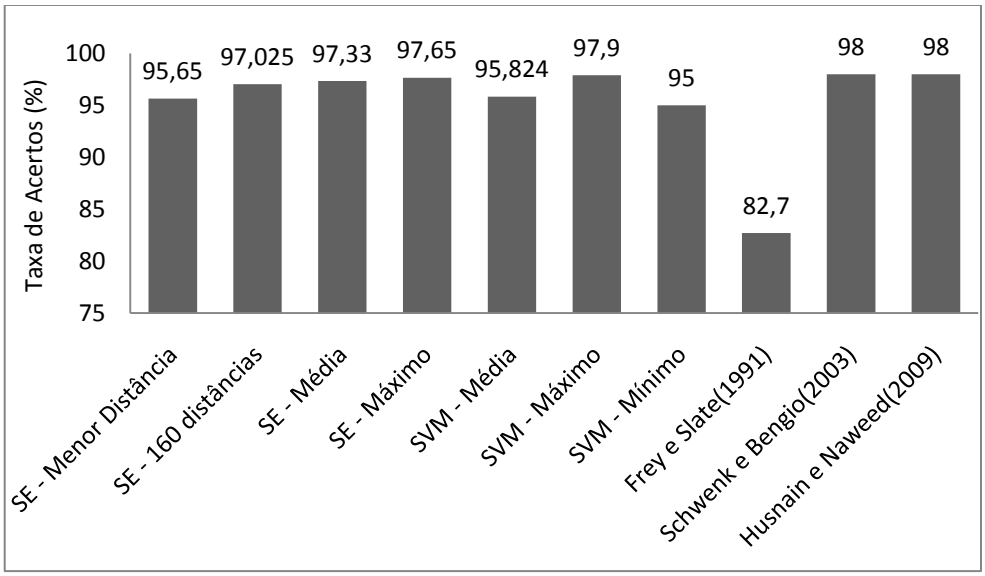

Figura 15: Resultados obtidos com este trabalho comparados aos resultados obtidos por Frey e Slate (1991), Schwenk e Bengio (2003) e Husnain e Naweed (2009)

Os sistemas propostos por este trabalho se comportaram de forma satisfatória em relação às taxas de acerto obtidas, comparando com comparados com as demais propostas apresentadas. O trabalho de Frey e Slate (1991) obteve uma taxa de acerto inferior aos demais trabalhos, possivelmente devido ao baixo poder tecnológico que existia na época da publicação deste trabalho. Schwenk e Bengio (2003) conseguiram taxa de acerto bastante elevada, porém deve ser considerado o alto custo computacional com treinamento de redes neurais que seu sistema possui. Husnain e Naweed (2009) também obtiveram resultados interessantes. Entretanto estes resultados foram obtidos selecionando

aleatoriamente 100 padrões dos 4000 que compõem a base de testes para a realização de experimentos.

Apesar de possuir resultados ligeiramente inferiores aos trabalhos de Schwenk e Bengio (2003) e Husnain e Naweed (2009), a utilização do sistema proposto por este trabalho se torna mais viável devido ao fato dos sistemas propostos possuírem alta eficácia na classificação dos caracteres mantendo um baixo custo computacional comparado com estes trabalhos.

6.5 Considerações finais

Neste capítulo foram apresentados os resultados obtidos pelos sistemas propostos. Foi apresentada também uma comparação com os trabalhos de Frey e Slate (1991), Schwenk e Bengio (2003) e Husnain e Naweed (2009). Comparado com estes trabalhos, ambos os sistemas propostos por este trabalho se comportaram de forma bastante satisfatória, atingindo uma alta taxa de acerto na classificação de caracteres e mantendo um baixo custo computacional. Outras propostas da literatura que trabalham com classificação de caracteres não foram contempladas na comparação de resultados devido à divergência na utilização e tratamento das bases de dados.

7
Conclusões

Este capítulo apresenta as conclusões desta monografia.

Esta monografia de conclusão de curso apresentou modelos para a realização de classificação de caracteres em imagens. Esta classificação pode ser utilizada para auxiliar a digitalização de documentos e seu posterior salvamento em formato digital sem que haja a necessidade de "re-escrever" todo o documento no computador.

Foram apresentados 2 modelos. O primeiro se baseia na técnica de Sistemas Especialistas, que é um método que possui um conjunto de regras do tipo "SE condição ENTÃO ação" baseados no calculo da distância euclidiana existente entre os padrões. Com base neste conjunto de regras é definido à que classe um determinado padrão de caractere pertence. O segundo método proposto se baseia na técnica de SVM voltadas para a classificação de padrões (SVC). Esta técnica pode ser descrita resumidamente como a geração de hiperplanos que separam as classes de caracteres de modo que se maximize a distância mínima entre padrões de classes diferentes.

Para a realização dos experimentos, foi utilizado um conjunto de dados obtido da *UCI Machine Learning Repository* (WEB1, 2009). Este conjunto de dados contém 20000 padrões contendo 16 atributos cada. Estes padrões que foram gerados aleatoriamente, levando em consideração vinte fontes diferentes do alfabeto romano e seis tipos de distorções. Vários outros conjuntos de dados foram gerados à fim de validar a eficácia do sistema.Com isto, pode-se comprovar que a configuração da base de dados, a distribuição dos padrões na base, influenciam diretamente na classificação.

Analisando outras propostas da literatura, os resultados obtidos por este trabalho foram satisfatórios. Ambos os modelos propostos obtiveram uma elevada taxa de acerto e um baixo custo computacional, tornando este sistema viável para a utilização. Pela análise dos resultados foi possível verificar a existência da relação direta entre o modo como a base de dados é configurada e os resultados obtidos pelos modelos propostos. Espera-se que estes modelos possam obter uma elevada taxa de acertos quando aplicado na classificação de outros tipos de caracteres, além dos caracteres do alfabeto inglês.

Como trabalho futuro pretende-se desenvolver um sistema que realize a extração de informações de caracteres através de uma imagem e defina os atributos necessários para representar o caractere reconhecido com o objetivo de gerar uma nova base de dados para ser utilizada em experimentos com os sistemas apresentados acima. Feito isto, pretende-se realizar a classificação de caracteres utilizando outras metodologias, além da realização de testes estatísticos para comparação dos resultados obtidos utilizando a base de dados apresentada neste trabalho e a base de dados a ser gerada pelo novo modelo.

8 Referências Bibliográficas

Aires, S.B.K., *Reconhecimento de caracteres manuscritos baseados em regiões perceptivas*. Dissertação de Mestrado. Pontifícia Universidade Católica do Paraná. Curitiba. 2005.

Assis, J.M.C. *Detecção de e-mails SPAM utilizando Redes Neurais Artificiais*. Dissertação de mestrado. Universidade Federal de Itajubá. 2006.

Baranoski, F.L., Justino, E.J.R. & Bortolozzi, F. *Identificação da Autoria em Documentos Manuscritos Usando SVM*. XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo – RS. 2005.

Batuwita, K.B. & Bandara, G.E. *Fuzzy Recognition of Offline Handwritten Numeric Characters*. ©IEEE. 2006.

Braga, I.L.S. *Identificação e Classificação de Litofácies com o uso da Teoria Bayesiana de Reconhecimento de Padrões*. Dissertação de mestrado. Universidade Estadual do Norte Fluminense – UENF. Macaé-RJ. 2005.

Campbell, J. B. *Introduction to remote sensing*. New York: The Guilford Press, 1996. p. 622

Carvalho, J.V., Sampaio, M.C. & Mongiovi, G. *Utilização de Técnicas de Data Mining para o Reconhecimento de Caracteres Manuscritos*. XIV Simpósio Brasileiro de Banco de Dados. Florianópolis. 1999.

Daugnam, J.G. *Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters*. J. Opt. Soc. Am. A, vol 2, No. 7. 1985.

Frey, P.W. & Slate, D.J. *Letter Recognition Using Holland-Style Adaptive Classifiers*. Machine Learning, 6, 161-182. Kluwer Academic Publishers, Boston. Manufactured in The Netherlands. 1991.

Freund, Y. & Schapire, R.E. *Experiments with a new boosting algorithm*. In *Machine Learning: Proceedings of Thirteenth International Conference.* p. 148–156, 1996.

Ganapathy, K., Fernando, C. G. & Davari, A.; Fast Character Recognition System Using Expert Systems. IEEE. 2005.

Holland, J.H.; *Escaping brittleness: The possibilities of general purpose machine learning algorithms applied to parallel rule-based systems*. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), Machine learning: An artificial intelligence approach (Vol. II). San Mateo, CA: Morgan Kaufmann Publishers. 1986.

Husnain, M. & Naweed, S. *English Letter Classification Using Bayesian Decision Theory and Feature Extraction Using Principal Component Analysis*. European Journal of Scientific Research. Vol.34 No.2. p.196-203. 2009

Jackson, P. *Introduction to Expert Systems. 3 ed. Harlow, England. Addison-Wesley*. 1998.

Kovács, Z.L. Redes Neurais Artificiais – Fundamentos e Aplicações. Editora Livraria da Física. Quarta Edição. 2002.

Lorena, A. C. & Carvalho, A. C. P. L. F. *Uma Introdução às Support Vector Machines*. Revista de Informática Teórica e Aplicada, Vol. 14, No 2. 2007.

Mancini, F., Yi, L.C., Pignatari,S.S.N.,Roque, A.C. & Pisa, I.T. *Aplicação de Redes Neurais Artificiais na Classificação de Padrões Posturais em Crianças Respiradoras Bucais e Nasais*. Revista de Informática Teórica e Aplicada. Volume XIV. Número 2. 2007

Máximo, A.O. & Fernandes, D. *Classificação supervisionada de imagens SAR do SIVAM pré-filtradas*. XII Simpósio Brasileiro de Sensoriamento Remoto. Goiânia. 2005

Motta, E. *Reusable Components for Knowledge Models.*Tese de doutorado. Knowledge Media University – Open University – UK. 1998.

Moussa, S.B., Zahour, A., Benabdelhafid, A. & Alimi, A.M. *New features using fractal multi-dimensions for generalized Arabic font recognition*. Pattern Recognition Letters. journal homepage: http://www.elsevier.com/locate/patrec .2009. Acessado em 14 de maio de 2010.

Nunes, C.M. *Seleção de primitivas utilizando algoritmo subida na encosta otimizado em problemas de reconhecimento de caracteres*. Dissertação de mestrado. Pontifícia Universidade Católica do Paraná. Curitiba. 2004.

Oliveira JR, J.J., Kapp, M.N., Freitas, C., Carvalho, J.M. & Sabourin, R. *Handwritten Month Word Recognition Using Multiple Classifiers*, XVII Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens – SIBGRAPI. Curitiba. 2004.

Rezende, S.O., *Sistemas Inteligentes – Fundamentos e aplicações.* Editora Manole LTDA. Primeira edição reimpressa. 2005

Rodrigues, R.J. & Thomé, A.C.G. Reconhecimento de dígitos cursivos – um método de segmentação por histogramas. VI Simpósio Brasileiro de Redes Neurais. Rio de Janeiro. 2000.

Santos, F.V.T. *Sistemas de Classificação.* Universidade do Minho. Escola de Engenharia. Departamento de sistemas de informação. 2000.

Semolini, R. *Support Vector Machines, Inferência Transdutiva e o Problema de Classificação*. Dissertação de mestrado. Universidade Federal de Campinas - UNICAMP. 2002.

Silva, A.M., Moita, G.F. & Almeida, P.E.M. *Um filtro anti-spam utilizando redes neurais artificiais Multilayer Perceptron*. XI Encontro de Modelagem Computacional. Volta Redonda – RJ. 2008.

Schwenk, H. & Bengio, Y. *Adaptive Boosting of Neural Networks for Character Recognition*. Dept. Informatique et Recherche Opérationnelle Université de Montréal, Montreal, Qc H3C-3J7, Canada. 1997.

Vapnik, V. *Statistical learning theory*. New York: Wiley. 1998

Velasques, E. *Classicação de Pontos de Segmentação de Dígitos Manuscritos.* Dissertação deMestrado. Pontifícia Universidade Católicado Paraná. Curitiba. 2006.

Veloso, L. R., *Reconhecimento de Caracteres Numéricos Manuscritos*: Dissertação de Mestrado, Universidade Federal da Paraíba, Campina Grande, 1998.

Vieira, A.C.H., Tedesco, P.C., Timóteo, A. & Lima, A. *Analisando Diálogos para Classificação de Padrões Utilizando Redes Neurais Artificiais e Árvores de Decisão.*XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo/RS. 2005.

Wang, W., Ding, X. & Liu, C.; *Optimized Gabor Filter Based Feature Extraction for Character Recognition*. State Key Laboratory of Intelligent Technology and Systems. Dept. of E.E.; Tsinghua Univ.; Beijing. 2002.

WEB 1; *Base de dados da UCI Learning Machine Repository.* http://archive.ics.uci.edu/ml/datasets/Letter+Recognition - Acessado em 14 de Outubro de 2009.

WEB 2; *Ferramenta LibSVM.* http://www.csie.ntu.edu.tw/~cjlin/libsvm - Acessado em 1 de maio de 2010.