DIEGO BERTOLINI GONÇALVES

AGRUPAMENTO DE CLASSIFICADORES NA VERIFICAÇÃO DE ASSINATURAS *Off-Line*

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Informática como requisito parcial para a obtenção do título de Mestre em Informática.

Curitiba PR
Outubro de 2008

DIEGO BERTOLINI GONÇALVES

AGRUPAMENTO DE CLASSIFICADORES NA VERIFICAÇÃO DE ASSINATURAS *Off-Line*

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Informática como requisito parcial para a obtenção do título de Mestre em Informática.

Área de concentração: *Ciência da Computação*

Orientador: Luiz Eduardo Soares de Oliveira
Co-orientador: Edson José Rodrigues Justino

Curitiba PR
Outubro de 2008

Gonçalves, Diego Bertolini

Agrupamento de Classificadores na Verificação de Assinaturas *Off-Line*. Curitiba, 2008. 86p.

Dissertação - Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática.

1. Verificação de Assinaturas *Off-Line* 2. Agrupamento de Classificadores 3. Algoritmos Genético. I. Pontifícia Universidade Católica do Paraná. Centro de Ciências Exatas e de Tecnologia. Programa de Pós-Graduação em Informática II-t.

Esta folha deve ser substituída pela ata de defesa devidamente assinada, que será fornecida pela secretaria do programa após a defesa.

Agradecimentos

A Deus.

Aos meus pais, Sérgio e Marlene.

Ao meu orientador Dr. Luiz Eduardo Soares de Oliveira, pela paciência, atenção, dicas e amizade.

As novas amizades que fiz, Cheila, Giovani, Priscila, Neimar e Eduardo, pela ajuda e companherismo.

As velhas amizadas, por estarem sempre perto.

A minha família pela força.

Aos professores do PPGIa pelo suporte, em especial ao Prof. e Co-orientador Dr. Edson José Rodrigues Justino.

Aos Professores da banca, Dr. George Darmiton da Cunha Cavalcanti e Dr. Jacques Facon, pelas contribuições para um trabalho melhor.

A todos que direta ou indiretamente contribuíram para realização desse trabalho.

Resumo

Neste trabalho apresentamos um estudo que visa reduzir o erro na identificação de falsificações em sistemas de verificação de assinaturas *off-line* através do agrupamento de classificadores (*ensembles*). Num total, quatro características (Distribuição de *Pixels*, Densidade de *Pixels*, Inclinação e Curvatura) e 16 diferentes configurações de *grids* são utilizados em nosso trabalho. O objetivo principal deste trabalho é formar agrupamentos de classificadores, através de características grafométricas e diferentes configurações de *grids*, melhorando assim o desempenho do sistema quanto à classificação, e, por conseguinte, reduzindo a falsa aceitação. Os agrupamento são formados através de um algoritmo genético clássico, onde três diferentes funções objetivos são propostas para avaliação. Dois diferentes cenários serão avaliados nesta pesquisa, no primeiro assumimos que só assinaturas genuínas e falsificações aleatórias são disponíveis. Já em outro, assinaturas genuínas, falsificações simples, aleatórias e simuladas são disponíveis durantes a formação dos agrupamentos. Avaliamos também o impacto que o número de assinaturas usadas como referência influem no desempenho do sistema. A base de dados utilizada nos testes é composta por 100 autores e avaliando os resultados pode-se afirmar que estes foram promissores.

Palavras-chave: Verificação de Assinaturas *Off-Line*, Agrupamento de Classificadores, Algoritmo Genético.

Abstract

In this work we discuss the use of ensemble of classifiers based on graphometric features to improve the reliability of the classification, hence reducing the false acceptance for signature verification systems. The ensemble was built using a standard genetic algorithm and different fitness functions were assessed to drive the search. Two diferent scenarios were considered in our experiments. In the former, we assume that only genuine signatures and random forgeries are available to guide the search. In the latter, on the other hand, we assume that simple and simulated forgeries also are available during the optimization of the ensemble. The pool of base classifiers are trained using only genuine signatures and random forgeries. Thorough experiments were conduct on a database composed of 100 writers and the results compare favorably.

Keywords: Off-Line Signature Verification, Ensemble of Classifiers, Genetic Algorithm.

Sumário

Resumo **ix**

Abstract **xi**

Lista de Figuras **xvi**

Lista de Tabelas **xvii**

Lista de Símbolos **xviii**

Lista de Abreviações **xix**

1 Introdução **1**
1.1 Descrição do Problema 2
1.2 Objetivos 4
1.3 Justificativas 4
1.4 Proposta 5
1.5 Contribuição 5
1.6 Organização 5

2 Revisão Bibliográfica **7**
2.1 Classificadores de Distância 7
2.2 Redes Neurais Artificiais 8
2.3 Cadeias Escondidas de Markov (HMMs) 9
2.4 Alinhamento Temporal Dinâmico 9
2.5 Máquinas de Vetores de Suporte 10
2.6 Técnicas Estruturais 10
2.7 Análise Crítica 11

3 Fundamentação Teórica **13**
3.1 Escritor-Independente e Dissimilaridade 13
3.2 Verificação de Assinaturas *Off-line* e *On-line* 14
3.3 Falsificações 16
3.4 Aprendizado de Máquina 17
3.4.1 Máquinas de Vetores de Suporte 18
3.5 Medidas de Desempenho 20
3.6 Curvas ROC 21

3.6.1 Área Abaixo da Curva ROC (AUC) 23
3.7 Esquemas de Fusão 24
3.7.1 Regra do Produto 26
3.7.2 Regra da Soma 27
3.7.3 Esquemas de Combinação de Classificadores 27
3.8 Algoritmos Genéticos 29
3.8.1 Componentes de um AG Clássico 31
3.9 Agrupamento de Classificadores (*Ensembles*) 33

4 Metodologia Proposta **35**
4.1 Definição do Problema 35
4.2 Definição da Base de Dados 35
4.2.1 Aquisição dos Dados 35
4.2.2 Segmentação 37
4.2.3 Dimensão dos Vetores de Características 38
4.3 Conjunto de Características 39
4.3.1 Distribuição de *Pixels* 39
4.3.2 Curvas de Bezier 40
4.3.3 Densidade de *Pixels* 41
4.3.4 Inclinação Axial 41
4.4 Classificação 42
4.5 Combinando Saídas dos Classificadores 43
4.6 Agrupamento de Classificadores (*Ensemble de Classificadores*) 45
4.7 Cenários Utilizados 46
4.8 Interpretação dos Resultados 46

5 Experimentos e Resultados **47**
5.1 Experimentos e Análise em Relação à Combinação da Saída dos Classificadores 47
5.2 Experimentos e Análise em Relação às Funções Objetivo 49
5.3 Experimentos e Análise em Relação ao Tamanho do Conjunto de Referências . 53
5.4 Análise quanto aos Classificadores Selecionados 57
5.5 Avaliação quanto aos Esquemas de Fusão usados com AGs 58

6 Conclusões **61**
6.1 Trabalhos Futuros 62

Lista de Figuras

1.1 Sistema automático de identificação. Adaptado de [Coetzer, 2005]. 2
1.2 Sistema genérico de verificação de assinaturas *off-line*. 2
1.3 Exemplos de assinatura: (a) Sobreposição de 3 assinaturas do mesmo autor demonstrando a variação intrapessoal, (b) e (c) Similaridade existente entre uma assinatura genuína e um falsificação. 3
1.4 Exemplos de assinatura: (a) genuína, (b) falsificação simples, e (c) falsificação simulada. 4

3.1 Arquitetura global da abordagem proposta. 14
3.2 Diagrama hierárquico quanto aos tipos de abordagens de vericação de assinaturas existentes. 15
3.3 Exemplo de assinaturas por região: (a) Assinatura Ocidental, (b) Assinatura Oriental, adaptado de [Ueda, 2003]. 16
3.4 Tipos de assinaturas ocidentais: (a) Assinatura Cursiva, (b) Rúbrica. 16
3.5 Exemplos de assinaturas: (a) Genuína; (b) Falsificação Aleatória; (c) Falsificação Simples; (d) Falsificação Simulada. 17
3.6 Tipos de falsificações. Adaptado de [Coetzer, 2005]. 17
3.7 Cenário onde hiperplanos separam os dados linearmente em duas classes. Os vetores de suporte encontram-se circulados. Adaptado de [Burges, 1998]. . . . 18
3.8 Intersecção existente entre assinaturas genuínas e falsificações. 21
3.9 Quatro situações possíveis em um classificador a partir de duas classes. 22
3.10 Gráfico ROC apresentando cinco classificadores discretos. Adaptado de [Fawcett, 2006]. 23
3.11 Típica curva ROC 24
3.12 Gráfico ROC. Área abaixo da curva (situação hipotética). Adaptado de [Fawcett, 2006]. 25
3.13 Exemplo típico de ótimo local e ótimo global. 30
3.14 Ciclo do Algoritmo Genético. 30
3.15 Exemplo de um *ensemble* formado por 3 classificadores distintos 33
3.16 Desempenho com o uso de *ensembles* onde as taxas de erros dos classificadores eram menores que 0.5 34

4.1 Metodologia proposta. 36
4.2 Exemplos de assinatura: (a) genuína, (b) falsificação simples, e (c) falsificação simulada. 37
4.3 Dois diferentes exemplos de configurações de *grids* usados para extração de características. 38

4.4 Dissimilaridades entre amostras genuínas do mesmo autor para gerar amostras positivas. A partir de quatro amostras genuínas, seis vetores de dissimilaridade são criados. 39
4.5 Dissimilaridade entre amostras genuínas de diferentes autores para gerar exemplos negativos. 39
4.6 Exemplo do método de Distribuição de *Pixels*. 40
4.7 (a) Assinatura genuína, e (b) Contornos da Assinatura. 40
4.8 (a) Exemplo de características extraídas do traçado e (b) exemplo de pontos detectados em um caso real, através da assinatura da Figura 4.7b. 41
4.9 Primitiva densidade de *pixels*. Adaptado de [Justino, 2001] 42
4.10 Primitiva inclinação axial. Adaptado de [Justino, 2001] 42
4.11 Ilustração do processo de extração da primitiva inclinação axial. Adaptado de [Santos, 2004] 43
4.12 Desempenho da base de classificadores. 44
4.13 Esquema de combinação das saídas dos classificadores utilizando um Sk = 5. . 45

5.1 Avaliação de desempenho quanto aos esquemas de combinação de classificadores. (a) *Sk* = 3 e (b) *Sk* = 15. 48
5.2 Comparação entre as três funções objetivos consideradas neste trabalho. *Sk* = 3. Cenário I. 50
5.3 Comparação entre as três funções objetivos consideradas neste trabalho. *Sk* = 9, Cenário I. 51
5.4 Comparação entre as três funções objetivos consideradas neste trabalho. *Sk* = 15, Cenário I. 51
5.5 Comparação entre as três funções objetivos consideradas nesse trabalho. *Sk* = 3, Cenário II. 52
5.6 Comparação entre as três funções objetivos consideradas nesse trabalho. *Sk* = 9, Cenário II. 52
5.7 Comparação entre as três funções objetivos consideradas nesse trabalho. *Sk* = 15, Cenário II. 53
5.8 Comparação entre diferentes números de (*Sk*) considerados nesse trabalho, usando como função objetivo a taxa de erro global, Cenário I. 55
5.9 Comparação entre diferentes números de (*Sk*) considerados nesse trabalho, usando como função objetivo a AUC, Cenário I. 56
5.10 Comparação entre diferentes números de (*Sk*) considerados nesse trabalho, usando como função objetivo a TPR fixada em 10%, Cenário I. 56
5.11 Classificadores selecionados para compor o agrupamento utilizando o Cenário I. Aptidão: (a) Erro Global, (b) AUC, (c) FPR fixada em 10%. 58
5.12 Classificadores selecionados para compor o agrupamento utilizando o Cenário II. Aptidão: (a) Erro Global, (b) AUC, (c) TPF fixada em 10%. 59
5.13 Classificadores selecionados para compor o agrupamento utilizando a AUC como aptidão e conjunto de validação: (a) Cenário II, (b) Cenário I 60

Lista de Tabelas

2.1 Bases de dados utilizadas na verificação de assinaturas. (256 N.C.: 256 Níveis de Cinza; I: Indivíduos; G: Genuínas; F: Falsificações; A: Amostras). 12

3.1 Métricas utilizadas em problemas com duas classes. 22

4.1 Variações para tamanhos de *grids* 38
4.2 Melhor classificador de cada conjunto de características referente ao conjunto de testes. 44

5.1 Avaliação do uso de diferentes esquemas de fusão para combinação das saídas de classificadores. 49
5.2 Taxa de erro global e AUC das diferentes funções de aptidão utilizadas, Cenário I. 49
5.3 Erro Global e AUC das diferentes funções objetivos utilizadas, Cenário II. . . . 50
5.4 Resultados dos experimentos utilizando a taxa de erro global como função objetivo, Cenário I. 54
5.5 Resultados dos experimentos utilizando a AUC como função objetivo, Cenário I. 54
5.6 Resultados dos experimentos utilizando o FPR fixada em 10% como função objetivo, Cenário I. 54
5.7 Resultados dos testes utilizando a taxa de erro global como função objetivo, Cenário II. 55
5.8 Resultados dos testes utilizando a AUC como função objetivo, Cenário II. . . . 57
5.9 Resultados dos testes utilizando a FPR fixada em 10% como função objetivo, Cenário II. 57

Lista de Símbolos

x	Objeto, padrão de entrada ou atributo.
$D(x, R)$	Vetor de dissimilaridade.
T	Conjunto de treinamento.
$f(x, y)$	Função discreta bidimensional.
Φ	Hiperplano de separação ótima.
C	Penalidade de erro no SVM.
ξ_i	Magnitude do erro de classificação.
α_i	Multiplicadores de Lagrange.
$K(s_i, x)$	Função do *kernel*.
$p(.\|.)$	Probabilidade *a posteriori*.
p_i	Objeto, primitivas.
^{p}	Grau do Polinômio.
σ	Desvio Padrão.
κ	Pode ser entendido como um fator (*scale*.)
δ	Deslocamento desejado.
e	Taxa de Erro Proporcionada por cada Classificador.
Q	Características Extraídas.
Z_i	Vetor de Dissimilaridades.

Lista de Abreviações

PPGIa	Programa de Pós-Graduação em Informática.
VIR	Laboratório de Visão, Imagem e Robótica.
PUCPR	Pontifícia Universidade Católica do Paraná.
AG	Algoritmo Genético.
AER	*Average Error Rate*.
EER	*Equal Error Rate*.
AUC	*Area Under Curve*.
FAR	*False Aceptance Rate*.
FRR	*False Rejection Rate*.
FP	*False Positive*.
TP	*True Positive*.
FN	*False Negative*.
TN	*True Negative*.
ROC	*Receiving Operator Characteristics*.
SVM	*Máquinas de Vetores de Suporte*.
HMM	Cadeias Escondidas de Markov.
RNA	Redes Neurais Artificiais.
DTW	Alinhamento Temporal Dinâmico.
k-NN	*k Nearest Neighbors*.
ESC	*Extended Shadow Code*.
RBP	*Resilient BackPropagation*
RBF	Redes com Funções de Base Radial.
DF	Característica de Direção.
TF	Característica de Transição.
R	Conjunto de Objetos de Referência.
Sq	Assinatura Questionada.
Sk	Assinatura de Referência.

Capítulo 1

Introdução

Sistema de verificação de assinaturas tem por objetivo verificar a autenticidade de uma assinatura através de métodos que possam discriminar uma assinatura genuína de uma falsificação [Batista et al., 2007]. De acordo com o método de aquisição da mesma, o processo de verificação de assinaturas pode ser classificado como *on-line* ou *off-line*. Para a abordagem *on-line* necessita-se de um *hardware* especial (mesa digitalizadora ou caneta sensível a pressão). Já na abordagem *off-line* tem-se a assinatura disposta em papel (cheque, contrato), sendo posteriormente digitalizada.

Atualmente, a verificação da identidade de pessoas é uma necessidade em todo o mundo, daí o crescimento do interesse em sistemas automáticos de identificação. Sistemas de identificação são sistemas utilizados para verificar ou reconhecer a identidade de pessoas. Assim, sistema automático de verificação de assinaturas *off-line* enquadram-se neste nicho. No entanto, segundo Plamondon e Shirari [Plamondon and Srihari, 2000], esse ocupa um nicho específico dentre os outros.

Ao estudar sistemas de identificação, esse variam bastante. De um lado, temos sistemas de identificação onde portamos um objeto que nos identifica (chaves, cartões, etc), do outro lado temos sistemas que necessitam de um conhecimento prévio (senhas, informações pessoais, etc). Sistemas de verificação de assinaturas também diferem dos sistemas biométricos baseados em propriedades do indivíduo (impressão digital, íris, retina, geometria da mão, face), pois nestes baseia-se em características fisiológias para a verificação do indivíduo e não no comportamento, como no caso da verificação de assinatura.

O diagrama proposto por Coetzer [Coetzer, 2005], ilustra onde sistema de verificação de assinaturas encontra-se dentro do campo de sistemas automáticos de identificação (Figura 1.1).

Verifica-se, no entanto, que mesmo a assinatura manuscrita não sendo o mais confiável meio de identificação, essa, é legalmente aceita e muitíssimo utilizada em diversos meios de transações.

A assinatura manuscrita, assim como a escrita, é um comportamento biométrico o qual é construído sobre um certo período de tempo na vida de um indivíduo. Características físicas e psicológicas influenciam vigorosamente na formação de tais comportamentos, além disso a assinatura de um indivíduo é socialmente e legalmente aceita como um firmamento de intenção sobre algo. Justino [Justino, 2001] em seu trabalho escreve:

> Uma assinatura constitui atualmente, no contexto jurídico, um dos meios para comprovar a intenção em transações envolvendo documentos.

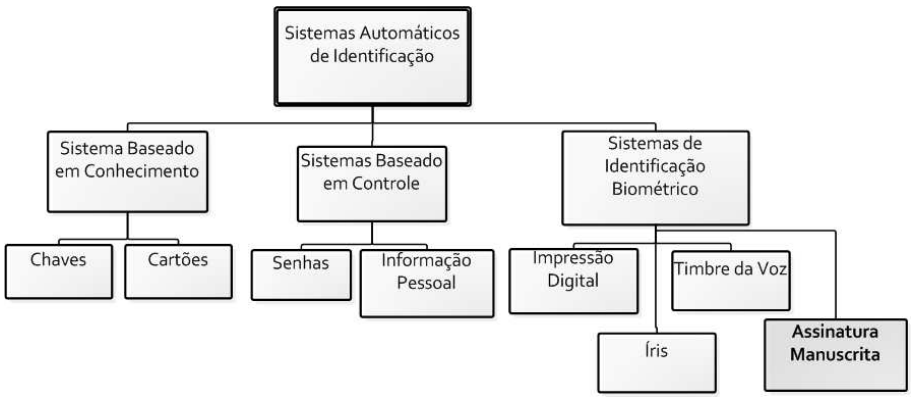


Figura 1.1: Sistema automático de identificação. Adaptado de [Coetzer, 2005].

Atualmente, muitos documentos necessitam ser assinados, como exemplo: cheques bancários, comprovantes de cartões de crédito, escrituras, etc. A partir disso, percebe-se a necessidade de sistemas de verificação automática de assinaturas.

Uma assinatura manuscrita em determinado documento caracteriza a intenção do autor envolvendo tal documento. Desse modo, quando um indivíduo assina um cheque bancário, contrato ou outro documento, esse passa a aceitar/concodar com o que consta no documento.

A Figura 1.2 apresenta um sistema genérico de verificação de assinaturas *off-line*.

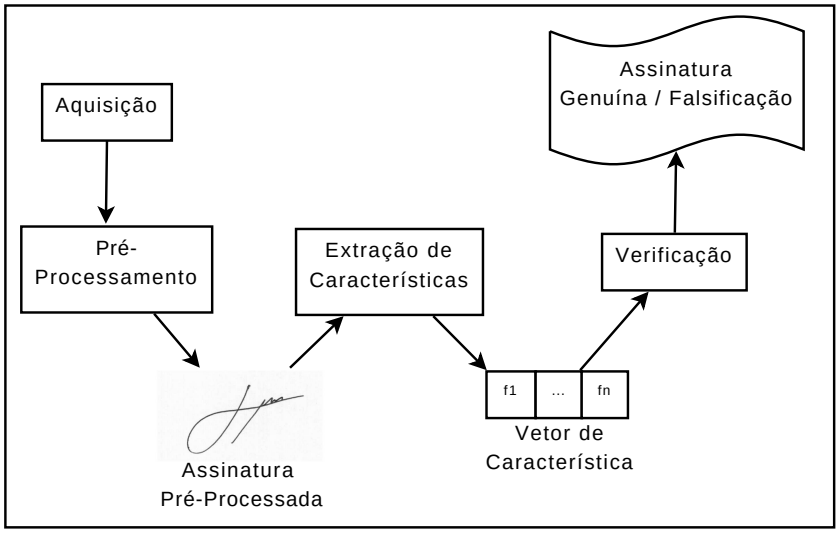


Figura 1.2: Sistema genérico de verificação de assinaturas *off-line*.

1.1 Descrição do Problema

Neste trabalho os métodos e estudos realizados valem-se para sistemas de verificação de assinaturas, tendo estes sentidos claros e diferentes de sistemas de reconhecimento de assinaturas. Sistemas de verificação de assinaturas consistem meramente em decidir se, dada uma assinatura em questão, e comparando-a a outras do mesmo autor, essa assinatura em questão pertence ou não pertence a esse escritor, ou seja, é uma assinatura genuína ou uma falsificação?

No caso de sistemas de reconhecimento de assinaturas, a idéia é: a partir de uma assinatura questionada, deseja-se saber quem é o autor, ou seja, quem assinou determinado documento, reconhecer o "proprietário" da assinatura.

Esta dissertação, baseia-se nos preceitos da área de reconhecimentos de padrões. Reconhecer padrões é uma característica nata do ser humano. Entretanto, esta tarefa não é tão simples de ser realizada computacionalmente. Mesmo em se tratando de duas classes somente, verificar assinaturas é tarefa árdua, tão complexa que observamos há décadas estudos sobre meios de automatizar tal processo, deixando ainda o problema em aberto.

Um dos fatores da verificação de assinaturas não ser trivial, se deve às fortes variações de características intrapessoal e a possíveis similaridades existentes entre falsificações e assinaturas genuínas. Verifica-se na Figura 1.3a que a variação intrapessoal e a similaridade entre uma falsificação e uma assinatura genuína pode ser alta (Figura 1.3b e 1.3c). Outro problema é que a assinatura de uma pessoa pode sofrer alterações ao longo dos anos, isso devido a uma série de fatores físicos e psicológicos intrínsecos a cada ser humano.

(a) (b) (c)

Figura 1.3: Exemplos de assinatura: (a) Sobreposição de 3 assinaturas do mesmo autor demonstrando a variação intrapessoal, (b) e (c) Similaridade existente entre uma assinatura genuína e um falsificação.

Coetzer *et al.* [Coetzer et al., 2006], apresentam um interessante estudo comparando o desempenho entre humanos e máquinas na verificação de assinaturas. O autor descreve em seu trabalho que os seres humanos apresentam altos índices de erros no processo de verificação de assinaturas. A partir disso pode-se perceber que mesmo sendo hábeis no processo de reconhecer padrões, afirmar com toda certeza, se dada uma assinatura em questão essa é genuína ou uma falsificação corresponde a uma tarefa árdua e geralmente atribuída a especialistas.

Os erros cometidos nos sistemas de verificação de assinaturas são classificados como erro tipo I e erro tipo II. O erro tipo I (falsa rejeição) ocorre quando o sistema classifica uma assinatura genuína como falsificação. Já o erro tipo II (falsa aceitação) ocorre quando o sistema classifica erroneamente uma falsificação como uma assinatura genuína. As falsificações geralmente são classificadas em três subconjuntos (aleatórias, simples e simuladas). A falsificação aleatória é normalmente uma amostra genuína de outro autor. A falsificação simples, ocorre quando o falsificador conhece o nome do autor, porém não possui um exemplo da assinatura a qual planeja falsificar. Por fim, a falsificação simulada ocorre quando o falsificador tendo posse de um ou mais exemplos de assinaturas genuínas, consegue fazer uma imitação da assinatura genuína [Coetzer, 2005]. A Figura 1.4, apresenta alguns exemplos de falsificações.

(a) (b) (c)

Figura 1.4: Exemplos de assinatura: (a) genuína, (b) falsificação simples, e (c) falsificação simulada.

1.2 Objetivos

Como objetivos principais deste trabalho pode-se citar:

- A redução dos erros tipo I e II em sistemas de verificação de assinaturas *off-line*;
- Avaliar o impacto do uso de agrupamento de classificadores baseado em características grafométricas.

Além disso, destacam-se os seguintes objetivos marginais:

- Avaliar dois possíveis cenários para agrupamento de classificadores. Em um primeiro momento assumimos que assinaturas genuínas, falsificações aleatórias, simples e simuladas são disponíveis para a construção do agrupamento. Já num segundo momento, para formação dos agrupamentos, assume-se possuir somente assinaturas genuínas e falsificações aleatórias (aplicações reais). Desta forma tem-se como avaliar o diferencial de desempenho ao possuir diferentes tipos de falsificações;
- Analisar o impacto do número de assinaturas de referência para o processo de treinamento;
- Analisar o desempenho de diferentes funções de aptidão durante a construção do agrupamento;
- Avaliar as características que apresentaram maior impacto na construção dos agrupamento.

A originalidade deste trabalho encontra-se fundamentada na avaliação de como o número de referências no processo de treinamento juntamente com o uso de diferentes tipos de falsificações podem impactar na taxa falsa aceitação utilizando agrupamento de classificadores.

1.3 Justificativas

Muitos trabalhos publicados recentemente focam em extrair primitivas relevantes de assinaturas. Entretanto, diferentes métodos são apresentados e muitas vezes as taxas ficam próximas umas das outras, [Armand et al., 2006], [Huang and Yan, 2002], [Fang et al., 2001].

A partir dessa observação, constata-se através de experimentos realizados que criar agrupamentos de classificadores através de regras de combinação utilizando Algoritmos Genéticos pode contribuir na otimização de resultados. Sendo que, não foram encontrados em literatura estudos abrangentes avaliando o desempenho do mesmo modo como realizamos neste trabalho.

1.4 Proposta

Para este trabalho de investigação, nossa proposta é minimizar as taxas de Falsa Rejeição e Falsa Aceitação em sistemas de verificação de assinaturas *off-line*. Para isto, utilizaremos diferentes classificadores, que agregados, através de diferentes regras possam apresentar uma melhor taxa de desempenho.

Combinar diversos classificadores utilizando diferentes regras e avaliar, para este conjunto de dados, qual deles apresenta um melhor desempenho. Como contamos com um grande número de classificadores, utilizaremos um método de busca (AGs) para explorar o grande espaço de busca em função da grande quantidade de classificadores.

1.5 Contribuição

Como contribuições científicas para este trabalho podemos destacar a análise crítica quanto ao processo de agrupamento de classificadores, avaliando sua importância para sistemas de verificação de assinaturas *off-line*. Através de estudos comparativos será possível avaliar o desempenho de algumas funções de aptidão, conseguindo assim, maximizar o desempenho de sistemas de verificação.

Outra contribuição importante se deve aos estudos em relação aos cenários utilizados, onde em um primeiro momento assumimos possuir assinaturas genuínas e falsificações simples, aleatórias e simuladas. Já num segundo momento, utilizamos somente assinaturas genuínas e falsificações aleatórias para avaliar o impacto causado no desempenho, o que é mais comum em aplicações comerciais. Assim que uma avaliação de desempenho quanto ao número de assinaturas de referência é realizada, consegue-se identificar um limiar para o processo de coleta de assinaturas genuínas de autores de acordo com o tipo de erro que o sistema deseja reduzir, pois sabe-se que na prática o número de assinaturas utilizadas para o processo de treinamento é, na maioria das vezes, limitado e pequeno (≤ 5).

1.6 Organização

Esta dissertação desenvolve-se ao longo de seis capítulos. Este capítulo contém uma breve descrição do trabalho proposto. O Capítulo 2 apresenta uma visão geral sobre o estado da arte relacionado a sistemas de verificação de assinaturas *off-line*. O Capítulo 3, apresenta um estudo sobre importantes trabalhos publicados ao longo dos anos, contribuindo com o leitor para um maior entendimento sobre técnicas e métodos computacionais utilizados nesta pesquisa. O Capítulo 4, apresenta, em detalhes, a metodologia proposta para o desenvolvimento deste trabalho. No Capítulo 5 são apresentados os resultados obtidos através dos experimentos realizados ao longo desta pesquisa. O último capítulo conclui o trabalho e indica trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

O objetivo deste capítulo é abordar os principais trabalhos relacionados ao tema desta dissertação. Existe um número considerável de trabalhos relacionados a verificação de assinaturas *off-line* conforme apresentam os recentes estudos de Batista *et al.* [Batista et al., 2007] e Impedovo [Impedovo and Pirlo, 2008].

Classifica-se os diversos trabalhos quanto à abordagem que cada autor utilizou para realizar a verificação. Assim teremos: Classificadores de Distância, Redes Neurais Artificiais, Cadeias Escondidas de Markov, Alinhamento Temporal Dinâmico, Máquinas de Vetores de Suporte e Técnicas Estruturais.

2.1 Classificadores de Distância

Nemcek e Lin [Nemcek and Lin, 1974], publicou um dos primeiros trabalhos sobre o assunto. Utilizando o método de máxima verossimilhança, conseguiram alcançar taxas de erros para tipo I e tipo II de 11% e 41%, respectivamente. Conforme descrito, as imagens de assinaturas estavam binarizadas e somente assinaturas genuínas e falsificações simples foram utilizadas. A base utilizada neste trabalho contava com 600 assinaturas genuínas produzidas por 15 autores e 120 falsificações simples cedidas por 4 indivíduos. O uso de falsificações simuladas em sistemas de verificação de assinaturas *off-line* é proposto por Ammar em [Ammar, 1991].

Diversos trabalhos passaram a utilizar assinaturas simples e simuladas, como Qi e Hunt [Qi and Hunt, 1994] que utilizou características globais e locais baseadas em *grids*, através de diferentes medidas de distâncias, alcançam taxas de erros para tipo I variando entre 3% e 11,3% e de 0% à 15% para erro do tipo II. Quinze autores produziram 300 assinaturas genuínas e 10 indivíduos colaboraram com 15 assinaturas cada. As imagens de assinaturas possuíam 256 níveis de cinza.

Utilizando classificadores de distâncias (*k-NN*), Sabourin e Genest [Sabourin and Genest, 1994] descrevem em seu trabalho alguns resultados onde propõem o uso do *Extended Shadow Code*. As assinaturas possuíam 256 níveis de cinza, sendo a base composta por 800 imagens de assinaturas onde 20 indivíduos produziram 40 assinaturas cada um. A média entre os erros de tipo I e II para experimentos usando *k-NN* foi de 0,01% para k = 1. Através dos classificadores de mínimas distâncias, a média dos erros foi de 0,77%. Em seus experimentos somente falsificações aleatórias foram utilizadas.

Fang *et al.* [Fang et al., 2001] desenvolveram um trabalho baseado no pressuposto de que os segmentos de assinaturas cursivas apresentam imperfeições, comparadas às assinaturas

genuínas. No processo de verificação é utilizado um classificador de distâncias. O método *leave-one-out* é utilizado para o treinamento e testes. A base utilizada é composta por 1320 assinaturas genuínas e 1320 falsificações, produzidas por 55 e 12 indivíduos respectivamente. As assinaturas encontravam-se em 256 níveis de cinza, em que somente falsificações simuladas foram avaliadas. Taxas de 18,1% para erro do tipo I e 16,4% para erro do tipo II foram obtidas.

2.2 Redes Neurais Artificiais

Segundo Batista *et al.*[Batista et al., 2007], o primeiro trabalho na área de verificação de assinaturas *off-line* a fazer uso de redes neurais foi proposto por Mighell *et al.* em [Mighell et al., 1989]. Utilizando um pequeno número de assinaturas e falsificações (80 genuínas e 66 falsificações) produzidas por um único indivíduo, realizando experimentos com imagens com 256 níveis de cinza e trabalhando apenas com falsificações simuladas o autor descreveu resultados para um EER de 2%.

O método proposto por Bajaj e Chaudhury [Bajaj and Chaudhury, 1997] utiliza características globais e locais relacionadas a assinatura. A partir de uma linha central principal da assinatura podemos encontrar os envelopes da assinatura, esses são na verdade curvas construídas com os pontos que se encontram acima ou abaixo da linha central. Desse modo, a curva situada acima da linha central é chamada envelope superior, enquanto a curva situada abaixo da linha central é conhecida por envelope inferior.

As imagens de assinaturas da base de dados usada por Bajaj e Chaudhury [Bajaj and Chaudhury, 1997] encontravam-se binarizadas. A base era composta por 150 assinaturas genuínas as quais foram produzidas por 10 autores e somente falsificações aleatórias foram utilizadas. Os resultados alcançados em seus experimentos apresentaram taxas de erros de 1% e 3 % para erros tipo I e II, respectivamente.

Guo *et al.* [Guo et al., 1997] utilizaram uma técnica que busca extrair características estáticas e dinâmicas das imagens de assinaturas. Assim, a técnica buscava segmentar a imagem através de pontos finais e junções. Contudo, através dos segmentos do traçado, podem-se extrair diversas características como: curvatura, centro de gravidade, comprimento, entre outras. Um ponto a ser revisto é que em assinaturas com alto grau de complexidade, um grande número de segmentos podem ser extraídos podendo estes não ter relevância para o processo de verificação.

Cardot *et al.* [Cardot et al., 1994] fez uso de abordagem global para detectar falsificações aleatórias. O uso de características como envelope e parâmetros geométricos (média da direção do traçado, momentos de inércia e escala). Através de uma base em que 300 indivíduos produziram um total de 6000 assinaturas genuínas e utilizando imagens com 256 níveis de cinza, os autores reportam taxas para erro tipo I e II de 5% e 2%, respectivamente.

Baltzakis e Papamarkos [Baltzakis and Papamarkos, 2001] utilizaram-se de características globais, *grid* de características e características de textura para representar cada assinatura. A fim de detectar falsificações aleatórias, o sistema proposto divide-se em duas etapas. Na primeira etapa, três redes MLPs (uma para cada conjunto de características) foram utilizadas e a Distância Euclidiana como uma métrica para uma primeira classificação. Uma Rede Neural RBF (*Radial Basis Function*) é treinada com as amostras que não foram usadas na primeira etapa, tomando uma decisão final. A base de dados foi composta por 2000 assinaturas (binarizadas) genuínas produzidas por 115 autores (15 à 20 assinaturas por autor). Das 2000 assinaturas, 1500 foram utilizadas no treinamento e 500 para testes. Foram utilizadas 57000 falsificações

aleatórias para os testes. Com relação aos resultados, taxas de 3% e 9,8% foram obtidas para erros do tipo I e II, respectivamente.

O método descrito a seguir, foi projetado originalmente a fim de reconhecer caracteres cursivos. Entretanto, Armand *et al.* [Armand et al., 2006], propõem o método para verificação de assinaturas *off-line*. Neste trabalho os autores, descrevem que o método emprega duas técnicas para extração de características: característica de direção (DF) e característica de transição (TF). A DF fundamenta-se na substituição da direção dos pixels no primeiro plano, em cinco possíveis direções: Vertical, Diagonal Direita, Horizontal, Diagonal Esquerda e Intercecção entre Linhas.

Com relação às características de transição, basicamente, essas armazenam informações sobre transições do primeiro para o segundo plano em uma imagem binária.

A base utilizada pelo autor consta de 2106 imagens de assinaturas, sendo essa composta por 39 grupos, em que cada grupo é composto por 24 assinaturas genuínas e 30 falsificações. Utilizaram-se redes neurais como classificador, sendo que dois algoritmos (RBP) e (RBF) foram avaliados. Somente falsificações simuladas foram consideradas. Em seus experimentos Armand *et al.* [Armand et al., 2006] obtiveram taxas de acertos de 91,21% para RPF e 88,00% para RBP.

2.3 Cadeias Escondidas de Markov (HMMs)

El-Yacoubi *et al.* [El-Yacoubi et al., 2000] utilizaram HMMs e princípios de validação cruzada para detectar falsificações aleatórias. Através de *grids* sobrepostos a imagem é computada a densidade de *pixels* existente em cada célula. A base de dados composta por 4000 assinaturas produzidas por 100 indivíduos, foi dividida em dois conjuntos, com 60 e 40 autores. Para ambos, foram usadas 20 assinaturas por autor para o treinamento e 20 para validação. Utilizando a regra do voto majoritário, chega-se a uma decisão final. A média entre os erros de tipo I e tipo II para as bases (60 / 40) são de 0,46% e 0,91%, respectivamente.

Justino *et al.* [Justino et al., 2001] em seu trabalho propõem a detecção de falsificações simples e simuladas. Utilizando-se de segmentação através de *grids* sobrepostos às imagens de assinaturas, Justino [Justino et al., 2001] extrai três características para cada célula do *grid* que são: densidade de *pixels*, distribuição de *pixels* e inclinação axial. Utilizando dois conjuntos de dados em seus experimentos, o primeiro sendo formado por 40 autores em que cada um produz 40 assinaturas genuínas e o segundo com 60 autores produzindo 40 assinaturas genuínas, 10 falsificações simples e 10 falsificações simuladas. O primeiro conjunto foi utilizado a fim de estabelecer o tamanho de um *codebook* para detecção de falsificações aleatórias. As taxas obtidas através das bases foram de 2,83% para erro do tipo I e para falsificações aleatória, simples e simulada (erro tipo II) foram obtidas 1,44%, 2,50% e 22,67%, respectivamente.

2.4 Alinhamento Temporal Dinâmico

Deng *et al.* [Deng et al., 1999] fez uso de duas bases, uma com assinaturas ocidentais (inglês) e outra para assinaturas orientais (chinês). Deng propõe o uso de um algoritmo de *Closed-Contour Tracing*, em que após o uso desse, os dados extraídos das curvas dos traçados são convertidos em sinais multiresolucional usando transformada Wavelet. Utilizou-se o DTW para a correspondência dos *zero-crossing* referentes as curvaturas dos dados. Resultados descritos pelo autor quanto à base ocidental é de 5,6% para erro do tipo I e 21,2% (simuladas) e

0% (simples) para erro do tipo II. Para o segundo conjunto de dados (chinês), as margens de erros foram de 6% para erro do tipo I e 13,5% para erro do tipo II (13,5% para simuladas e 0% para simples). Um total de 3500 assinaturas genuínas ocidentais e orientais foram utilizadas, produzidas por 100 indivíduos. As imagens de assinaturas encontravam-se em 256 níveis de cinza e falsificações simples e simuladas são utilizadas.

Fang *et al.* [Fang et al., 2003] a fim de lidar com variações intrapessoais propõem uma abordagem baseada em DTW e uma projeção de contornos unidimensional. Com o objetivo de detectar falsificações simuladas, um DTW não linear é aplicado porém, de maneira diferente. Em seu método, Fang *et al.* [Fang et al., 2003], ao invés de utilizar a distância entre uma assinatura genuína e uma amostra de referência para tomada de decisão, utiliza uma distorção posicional de cada ponto da projeção do contorno, incorporada em uma medida de distância. Através de método de validação cruzada *leave-one-out* e distância de Mahalanobis taxas de erro médio (AERs) de 20,8% e 18,1% foram obtidas através de experimentos com imagens binárias e imagens com 256 níveis de cinza, respectivamente. A base é formada por 1320 assinaturas genuínas produzidas por 55 autores e o mesmo número para falsificações, porém produzidas por 12 autores.

2.5 Máquinas de Vetores de Suporte

Um interessante trabalho de comparação de desempenho de classificadores é feito por Justino *et al.* [Justino et al., 2005]. Neste trabalho Justino avaliou o desempenho entre SVM e HMM na detecção de falsificações aleatórias, simples e simuladas. Foram utilizados *grids* sobrepostos à imagem para processo de segmentação, assim, características estáticas e pseudo-dinâmicas são utilizadas. O autor utilizou a Densidade de *Pixels*, Distribuição de *Pixels*, Curvatura dos Ângulos e a Inclinação como características estatícas e pseudo-dinâmicas. Utilizando SVM com um *kernel* linear obteve melhores resultados que com HMM.

Outro estudo comparativo é realizado por Ozgunduz *et al.* [Ozgunduz et al., 2005], em que realiza experimentos com o classificador SVM e Redes Neurais Artificiais. Utilizou-se características geométricas globais como, direção e *grids* de características na representação de assinaturas. Através de um *kernel* RBF para o SVM e o algoritmo *Backpropagation* para o treinamento da RNA, foram obtidos erros do tipo I e II para o SVM de 0,02% e 0,11% e de 0,22% e 0,16% para a RNA. Ozgundus *et al.* [Ozgunduz et al., 2005] para este trabalho utiliza 1320 exemplos (não especificado) produzidos por 70 indivíduos, com objetivo de detectar falsificações aleatórias e simuladas. Para os dois casos foram utilizadas falsificações simuladas para o treinamento dos classificadores.

2.6 Técnicas Estruturais

Huang e Yan [Huang and Yan, 2002] apresentaram um sistema baseado em duas etapas: RNA e Técnicas Estruturais. Características direcionais de bordas e geométricas foram utilizadas para representar as assinaturas. Na primeira etapa, atribui à assinatura três possíveis classificações: liberada (assinaturas genuínas), reprovada (falsificação simulada mal reproduzida ou aleatória) e questionável (falsificações simuladas). Para essas assinaturas questionáveis, utilizou-se na segunda fase o algoritmo a fim de verificar características estruturais, para comparar a correlação detalhada da estrutura entre as assinaturas de teste e os exemplos de referência.

Em seus experimentos a rede neural rejeitou 2,2% das assinaturas genuínas, aceitou 3,6% das falsificações e ficou indecisa em 32,7%. O segundo classificador rejeitou 31,2% das assinaturas genuínas questionadas e aceitou 23,2% das falsificações questionáveis. Combinando os classificadores, uma taxa de 6.3% foi alcançada para erro do tipo I e de 8.2% para erro do tipo II. A base utilizada por Huang e Yan [Huang and Yan, 2002] é composta de 1272 assinaturas genuínas produzidas por 53 autores e 7632 falsificações produzidas por 53 indivíduos.

Ismail e Gad [Ismail and Gad, 2000], a fim de verificar assinaturas árabes, utilizaram conceito Fuzzy e características locais como: linha central, ângulo de curvas, ângulos de linhas, pontos de funcionalidades críticas e círculos centrais. Ao invés de utilizar-se de um limiar para decisão, um conjunto de regras fuzzy é utilizado para tomada de decisão com um grau de certeza. Através de uma base composta por 22 autores, na qual seis assinaturas usadas para treinamento, quatro assinaturas genuínas e 5 falsificações simuladas para testes (por autor). Os autores, apresentaram uma média de erro global de 98% para estes experimentos.

2.7 Análise Crítica

Concluindo, o estudo de alguns trabalhos citados neste capítulo busca contribuir para elaboração desta pesquisa, ajudando a entender a complexidade deste projeto e observar os resultados alcançados através de cada método. Todavia, um estudo comparativo entre resultados obtidos considerando as abordagens utilizados para o processo de verificação torna-se difícil devido à diversidade de bases de dados existentes. Avaliando resultados, temos as vezes baixas taxas de erros, contudo ao realizar um estudo detalhado, percebemos que tratam apenas de alguns tipos de falsificações, em que geralmente desconsideram falsificações simuladas. A Tabela 2.1 apresenta as principais características das bases utlizadas em cada trabalho. No capítulo seguinte, apresentamos detalhadamente a Fundamentação Teórica utilizada neste trabalho.

Tabela 2.1: Bases de dados utilizadas na verificação de assinaturas. (256 N.C.: 256 Níveis de Cinza; I: Indivíduos; G: Genuínas; F: Falsificações; A: Amostras).

Referências	Imagens	Assinaturas	Tipos de Falsificações
[Nemcek and Lin, 1974]	Binárias	600G / 15I - 120F / 4I	Simples
[Qi and Hunt, 1994]	256 N.C.	300B/15I - 150F /10I	Simples e Simulada
[Sabourin and Genest, 1994]	256 N.C.	800G /20I	Aleatórias
[Fang et al., 2001]	256 N.C.	1320G/55I - 1320F/12I	Simuladas
[Mighell et al., 1989]	256 N.C.	80G/1I - 66F/1I	Simuladas
[Bajaj and Chaudhury, 1997]	Binárias	150G/10I	Aleatórias
[Cardot et al., 1994]	256 N.C.	6000G/300I	Aleatórias
[Baltzakis and Papamarkos, 2001]	Binárias	2000G/115I	Aleatórias
[Armand et al., 2006]		936G/39I - 1170F/39I	Simuladas
[El-Yacoubi et al., 2000]	Binárias	4000G/100I	Aleatórias
[Justino et al., 2001]	256 N.C.	4000G/100I - 1200F/10I	Simples e Simulada
[Deng et al., 1999]	256 N.C.	1000G/50I - 2500G/50I	Simples e Simulada
[Fang et al., 2003]	256 N.C.	1320G/55I - 1320F/12I	Simuladas
[Justino et al., 2005]	256 N.C.	4000G/100I - 1200F/10I	Simples e Simulada
[Ozgunduz et al., 2005]	256 N.C.	1320A/70I	Simuladas
[Huang and Yan, 2002]	256 N.C.	1272G/53I - 7632F/53I	Simuladas

Capítulo 3

Fundamentação Teórica

Este capítulo apresenta as técnicas computacionais utilizadas em nosso trabalho, contribuindo assim com uma base de entendimento dos métodos computacionais utilizados e como estes funcionam. Entretanto, uma maior riqueza de detalhes pode ser encontrada nas referências bibliográficas citadas aqui. De início, apresentaremos alguns conceitos-chave da área de verificação de assinaturas *off-line*.

3.1 Escritor-Independente e Dissimilaridade

Srihari *et al.* [Srihari et al., 2004], apresentam uma categorização sobre métodos de verificação de assinaturas, como escritor-dependente e independente.

No processo de verificação de escritor-independente, tem-se o interesse em classificar uma assinatura em termos de sua autenticidade (verdadeira ou falsa), dessa forma pode-se reduzir qualquer problema envolvendo reconhecimento de padrões em duas classes.

A abordagem aqui utilizada é a mesma utilizada por peritos forenses, que comparam exemplos de assinaturas questionadas (Sq) com algumas amostras de assinaturas de referência (Sk), conseguindo, assim, afirmar se: Dada um assinatura questionada (Sq) em comparação a algumas amostras de referência (Sk) esta é uma assinatura genuína ou uma falsificação? Para este processo de comparação são extraídas diferentes características para computar o grau de semelhança entre os exemplos disponíveis.

Conceitos de similaridade, dissimilaridade, e proximidade são discutidos em literatura em diferentes perspectivas [Srihari et al., 2004], [Oliveira et al., 2007], [Santini and Jain, 1999]. Pekalska e Duin [Pekalska and Duin, 2002], apresentam em seu trabalho a idéia de representar a relação entre objetos através da dissimilaridade, chamando de representação de dissimilaridade. Desta forma, cada objeto é representado através da diferença de um conjunto de objetos de referência, chamados de conjunto de representação R. Cada objeto x é representado por um vetor de dissimilaridade $D(x, R) = [d(x, p_1), d(x, p_2), ..., d(x, p_n)]$ para os objetos $P_i \in R$.

Seja R representado por um conjunto composto por n objetos. Um conjunto de treinamento T composto por m objetos é representado por uma matriz $m \times n$ sendo a dissimilaridade $D(T, R)$. Neste contexto, observa-se que a forma de classificar um novo objeto x, é representada por $D(x, R)$ utilizando seu vizinho mais próximo. O objeto x a ser classificado é classificado na classe de seu vizinho mais próximo, ou seja, a classe de representação do objeto p_i dado por $d(x, p_i) = min_{p \in R} D(x, R)$.

Em outra abordagem, cada dimensão corresponde a uma dissimilaridade $D(.,p_i)$ para um objeto p_i, assim, as dimensões transportam informações de tipo homogêneo. Para tal, as diferenças entre objetos semelhantes devem ser pequenas (objetos que pertençam à mesma classe) e grande para objetos de classes diferentes. Dessa forma, $D(.,p_i)$ pode ser interpretado como um atributo.

Um conceito relacionado é o de vetor de dissimilaridade, nesse caso a idéia consiste em extrair vetores de características das assinaturas questionadas (Sq) e assinaturas de referência (Sk), calculando o vetor de dissimilaridade de características. Assim, para as assinaturas provenientes do mesmo autor (genuínas) todos os componentes do vetor de dissimilaridade devem ser próximo de 0, caso sendo uma falsificação, os componentes devem ser bem maiores que 0.

Nesse caso, utiliza-se um conjunto de referência composto por n exemplos de assinaturas genuínas Sk_i, $i = 1, 2, 3, ..., n$. Comparando, então, cada Sk com um exemplo de amostra questionada Sq. Seja V_i as características extraídas dos exemplos de assinaturas de referências (Sk) e Q as características extraídas das assinaturas questionadas (Sq), então, o vetor de dissimilaridade de características $Z_i = |V_i - Q|$ é computado para alimentar os classificadores C_i que proporcionam uma decisão parcial, sendo a decisão final D dada através de esquemas que combinam as saídas dos classificadores obtendo um consenso na decisão (geralmente a regra do voto majoritário é utilizada). A Figura 3.1 representa o método baseado em vetor de dissimilaridade.

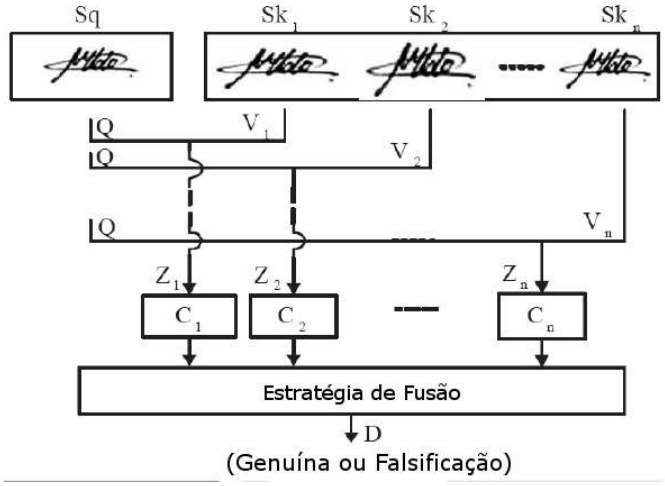


Figura 3.1: Arquitetura global da abordagem proposta.

3.2 Verificação de Assinaturas *Off-line* e *On-line*

A diferença entre os dois métodos de verificação de assinaturas (*Off-line* e *On-line*) dá-se pelo mecanismo de aquisição dos dados. Tem-se, atualmente, meios de obter o sinal referente a assinatura do autor de modo automático e em tempo de execução, conforme o autor assina em um equipamento especial, sua assinatura é digitalizada e diversas características são capturadas em tempo real. Esse método é chamado de *on-line* (ou dinâmico). Entretanto, em muitos casos o uso de equipamentos especiais para captura da imagem não há como ser usado, como por exemplo: cheques, escrituras, comprovante de cartão de crédito, entre outros. Necessita-se,

então, digitalizá-los após serem previamente assinadas em papel. Tal método é conhecido como *off-line* (ou estático).

O método *on-line* apresenta diversas vantagens se comparado ao *off-line* [Plamondon and Srihari, 2000]. Contudo, devido à necessidade de equipamentos especiais para captura da informação, esse é ainda um método menos utilizado. No método *on-line* consegue-se capturar características dinâmicas da assinatura durante os movimentos realizados ao longo do documento [Plamondon and Lorette, 1989].

Já na abordagem *off-line* após a digitalização temos uma imagem digital, a qual pode ser considerada uma função discreta bidimensional $f(x, y)$. Segundo Justino [Justino, 2001] um ponto importante do método *off-line* se refere à capacidade de obter dados mais pertinentes ao autor da assinatura, contribuindo na viabilidade do processo de suplementação das características usadas.

Esse método apresenta uma série de características que o torna mais desafiador abordando diferentes áreas da computação. O diagrama hierárquico da Figura 3.2, apresenta a estrutura de classificação dos métodos de verificação de assinaturas.

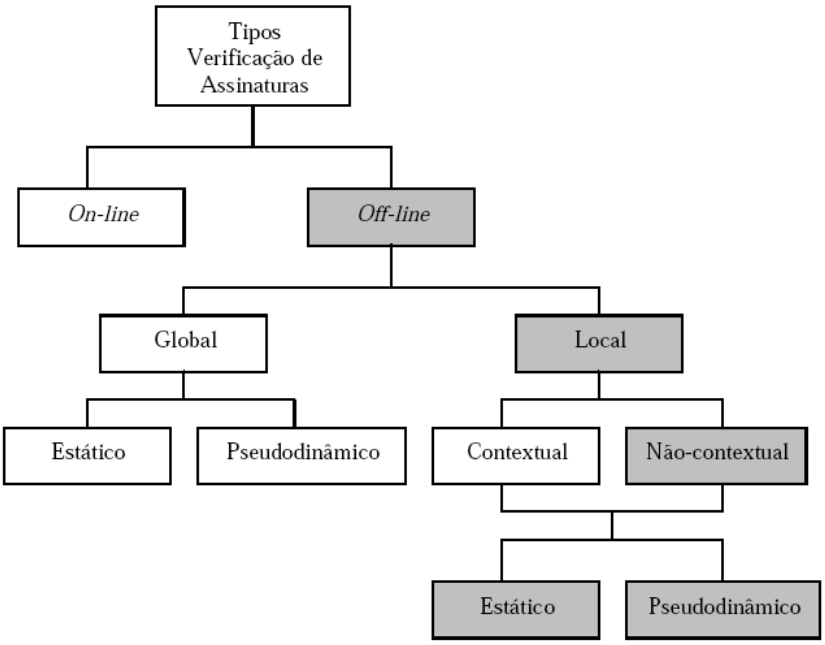


Figura 3.2: Diagrama hierárquico quanto aos tipos de abordagens de vericação de assinaturas existentes.

- Estática: Capacidade de representação de características relacionadas com a forma da imagem da assinatura, como por exemplo a altura e o comprimento.
- Pseudo-Dinâmica: Capacidade de representação de características relacionadas à dinâmica da escrita, assim como a inclinação e a curvatura.

3.3 Falsificações

A assinatura está fortemente ligada à forma da escrita da região de origem do autor e claro com o alfabeto por ele utilizado (Ocidental / Oriental), conforme mostra a Figura 3.3.

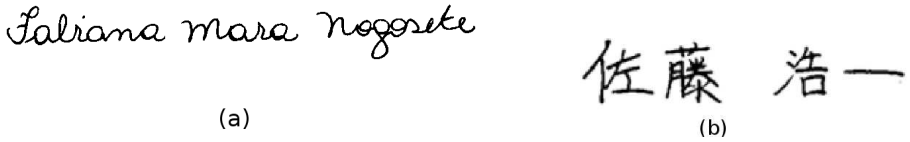


Figura 3.3: Exemplo de assinaturas por região: (a) Assinatura Ocidental, (b) Assinatura Oriental, adaptado de [Ueda, 2003].

Classificam-se as assinaturas ocidentais como sendo cursivas ou rubricas. No estilo de assinatura cursiva, o autor assina escrevendo o próprio nome. O modo cursivo advém da forma de escrita manuscrita a qual estamos acostumados. Já a rubrica, apresenta padrões complexos, em que dificilmente consegue-se reconhecer e interpretar caracteres presentes. Não há regra para uma rubrica, o autor pode utilizar caracteres, formas ou desenhos estilizados por ele. A Figura 3.4 apresenta os dois estilos de assinaturas ocidentais.

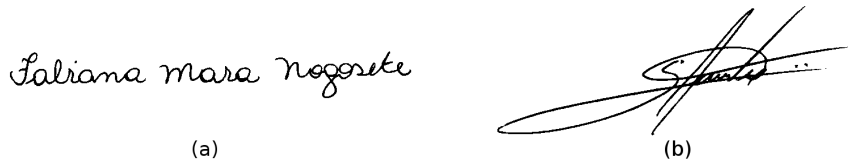


Figura 3.4: Tipos de assinaturas ocidentais: (a) Assinatura Cursiva, (b) Rúbrica.

De acordo com Coetzer [Coetzer, 2005] pode-se classificar as falsificações em três tipos: aleatória, simples e simulada. Vejamos uma breve descrição sobre estes três possíveis tipos de falsificações.

- **Falsificação Simulada**: A falsificação simulada conhecida também como hábil, é reproduzida pelo falsificador quando esse detém em seu poder um ou mais modelos da assinatura genuína do autor, na qual, através do modelo de referência, o falsificador tenta copiar com exatidão a assinatura verdadeira.

- **Falsificação Simples**: Nesse tipo, o falsificador conhecendo o nome do autor à qual deseja-se falsificar assinatura, apenas escreve-o de maneira manuscrita, não incluindo características pertinentes ao autor. O fato é que a falsificação pode ou não ter similaridade com a genuína.

- **Falsificação Aleatória**: O falsificador cria uma assinatura para o autor sem um conhecimento da assinatura genuína. Com isso na maioria dos casos a falsificação não possui semelhança com a original. Na prática utiliza-se a assinatura de outro autor para teste com falsificações aleatórias.

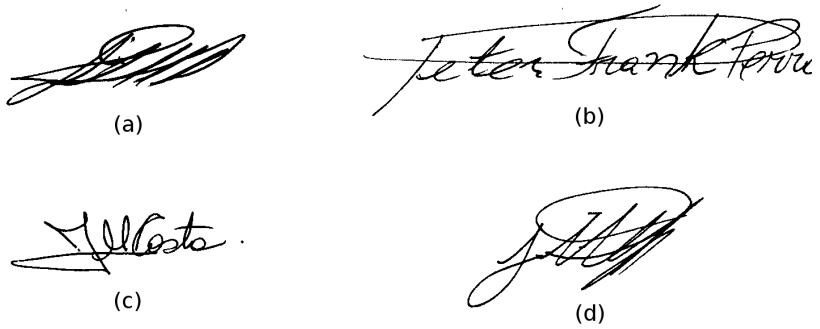


Figura 3.5: Exemplos de assinaturas: (a) Genuína; (b) Falsificação Aleatória; (c) Falsificação Simples; (d) Falsificação Simulada.

A Figura 3.5 apresenta exemplos dos tipos de falsificações mencionados.

Alguns autores como Coetzer [Coetzer, 2005] e Kalera [Kalera et al., 2004], consideram a habilidade do indivíduo que reproduziu a assinatura, classificando este como profissional ou amador. O esquema apresentado na Figura 3.6 demonstra tais classificações.

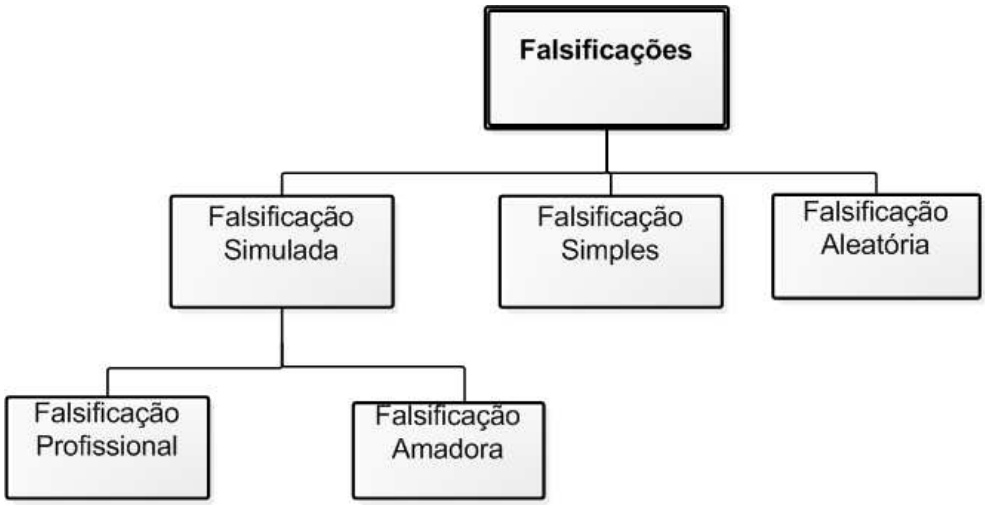


Figura 3.6: Tipos de falsificações. Adaptado de [Coetzer, 2005].

Nesta pesquisa, compromete-se a estudar e avaliar a taxa de acertos/erros sobre os três tipos de falsificações. Pois alguns trabalhos avaliam o desempenho perante um ou outro tipo de falsificação, [Kalera et al., 2004], [Coetzer, 2005].

3.4 Aprendizado de Máquina

O processo de aprendizagem indutiva pode ser classificado como *supervisionado* e *não supervisionado*. No aprendizado supervisionado, um conjunto de exemplos rotulados é fornecido ao algoritmo. No aprendizado não supervisionado, o algoritmo de aprendizado realiza uma análise dos exemplos fornecidos e tenta, de algum modo, agrupar estes exemplos com base em algo que sejam inerentes aos mesmos [Duda et al., 2000]. Neste trabalho utiliza-se o algoritmo

de aprendizagem supervisionado SVM (*Suport Vector Machine*). Através dos estudos de Justino *et al.* [Justino et al., 2005], avaliando o desempenho de diferentes métodos de aprendizado, demonstra a eficiência do SVM com problemas envolvendo 2 classes. Apresenta-se a seguir uma visão geral sobre o SVM, embasado nos trabalhos de Vapnik [Vapnik, 1995].

3.4.1 Máquinas de Vetores de Suporte

Máquinas de vetores de suporte (SVM) consiste em uma técnica para o treinamento de classificadores baseando-se na minimização de risco estrutural [Burges, 1998]. Proposto por Vladimir Vapnik [Vapnik, 1995], o método é essencialmente uma abordagem geométrica para o problema de classificação. Atualmente, vem sendo largamente utilizado em problemas de reconhecimento de padrões. Neste caso, vemos que a complexidade da hipótese é relativa à margem que os dados são separados e não ao número de atributos.

Considere a separação dos dados em duas classes, isto através de um hiperplano de separação, realizado pelo classificador SVM. Tem-se que, um hiperplano é ótimo quando este separa os dados com a máxima margem possível, sendo definida pela soma dos pontos positivos e negativos mais próximos do hiperplano. Tais pontos são conhecidos como vetores de suporte. Na Figura 3.7, os vetores de suporte encontram-se circulados.

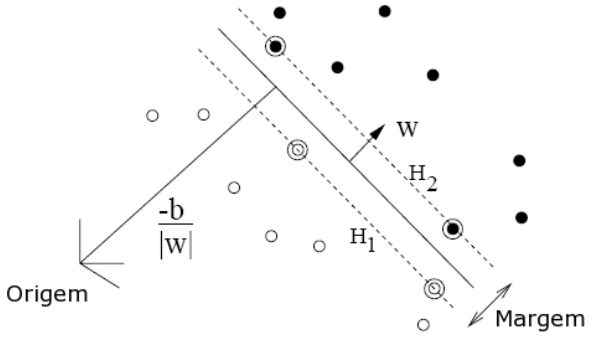


Figura 3.7: Cenário onde hiperplanos separam os dados linearmente em duas classes. Os vetores de suporte encontram-se circulados. Adaptado de [Burges, 1998].

A partir de um conjunto de treinamento $\{x_i, y_i\}, y_i \in \{-1, 1\}, x_i \in \Re^d$, onde x_i representa o enésimo elemento de entrada e y_i representa um rótulo de classe para $x_i, i = 1, ..., l$. Para o cálculo do hiperplano num problema de classificação binária (2 classes) buscando uma margem ótima no cálculo do hiperplano, esse hiperplano é definido pela equação $x \cdot w + b = 0$, onde w é a normal ao hiperplano. Pode-se assumir essa equação como a equação de uma reta na forma $z = a \cdot x + b$. Sendo que para todo vetor x que faça parte deste hiperplano este deve satisfazer a equação, onde w e b correspondem respectivamente a inclinação e deslocamento da reta. Desta forma, a base de treinamento é dividida da seguinte forma:

$$x_i \cdot w + b \geq +1 \; para \; y_i = +1 \tag{3.1}$$

$$x_i \cdot w + b \leq -1 \; para \; y_i = -1 \tag{3.2}$$

Podendo estas ser combinadas na inequação:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \; \forall i \tag{3.3}$$

Todavia, nas aplicações reais um conjunto de dados dificilmente é separável através de um hiperplano linear. Para tal adicionam-se variáveis de alargamento de margem ξ_i, "relaxando" as restrições do SVM linear. Com isso permitimos algumas falhas na margem. São computadas essas falhas penalizando-as através de uma variável de controle. Desta forma, passa-se a ter:

$$x_i \cdot w + b \geq +1 - \xi_i \; para \; y_i = +1 \tag{3.4}$$

$$x_i \cdot w + b \leq -1 + \xi_i \; para \; y_i = -1 \tag{3.5}$$

Minimizando $\| \; w \; \|^2$ em função da Equação 3.3, encontra-se o hiperplano com margem ótima. Contudo, esse é um problema quadrático de otimização, em que utiliza-se Multiplicadores de Lagrange α_i. Adicionando coeficientes de Lagrange positivos para cada restrição presente na Equação 3.3, sendo esse multiplicado pela restrição 3.3 e subtraídos da função objetivo ($\| \; w \; \|^2$), resultando em:

$$L_P \equiv \frac{1}{2} \| \; w \; \|^2 - \sum_i \alpha_i y_i (x_i \cdot w + b) + \sum_i \alpha_i \tag{3.6}$$

A princípio deve-se minimizar L_P em relação a w e *b* e simultaneamente derivar L_P em relação a α_i tendendo a zero, sujeito $\alpha_i \geq 0$ (restrição *R*1). O segundo passo é maximizar L_P sujeito a restrição que o gradiente de L_P tende a zero em relação a w e *b*, em que $\alpha_i \geq 0$ (restrição *R*2). Isso é chamado de problema *"dual"*, onde o mínimo de L_P para w e *b* e α (*R*1) são os mesmos valores para o máximo de L_P para (*R*2). Assim, quando gradientes de L_P tendem a zero, obtém-se:

$$w = \sum_i \alpha_i y_i x_i \tag{3.7}$$

$$\sum_i \alpha_i y_i = 0 \tag{3.8}$$

onde substituindo as equações (3.7) e (3.8) em (3.6), temos:

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{3.9}$$

devendo ser maximizada, conforme:

$$0 \leq \alpha_i \leq C_i \tag{3.10}$$

$$\sum_i \alpha_i y_i = 0 \tag{3.11}$$

sendo C_i a tolerância de erros ao hiperplano.

O número total de vetores de suporte é representado por *Ns*. Contudo, isto é aplicado a dados que são linearmente separáveis. Assim sendo a solução de vetores para estes hiperplanos é dada por:

$$\sum_{i=1}^{Ns} \alpha_i y_i x_i \tag{3.12}$$

Como na prática dificilmente se possui dados linearmente-separáveis, às vezes existe a necessidade de mapear o espaço de entrada $\Re^d$ para um outro espaço de dimensão mais alta (H). Isso equivale a "distorcer" o espaço geométrico ou inserir novas dimensões.

$$\Phi : \Re^d \mapsto H \tag{3.13}$$

Pode-se utilizar a Equação (3.14) para fazer o mapeamento.

$$K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \tag{3.14}$$

Através da equação 3.15 pode-se encontrar o lado do hiperplano que um vetor *x* esta.

$$\sum_{i=1}^{Ns} \alpha_i y_i K(s_i, x) + b \tag{3.15}$$

A função $K(s_i, x)$ é conhecida como Funções de Núcleo ou (*Kernel Functions*), s_i nesse caso é um vetor de suporte e x é um vetor de teste.

A literatura apresenta diversos *kernels* utilizados com sucesso em problemas de reconhecimento de padrões [Burges, 1998]. Os *kernels* $K(s_i, x)$ mais conhecidos são: Polinomial, Gaussiano e Sigmóidal, como descritos respectivamente:

$$K(x, y) = (x \cdot y + 1)^p \tag{3.16}$$

$$K(x, y) = e^{-\|x-y\|^2/2\sigma^2} \tag{3.17}$$

$$K(x, y) = tanh(\kappa x \cdot y - \delta) \tag{3.18}$$

3.5 Medidas de Desempenho

A análise de desempenho em sistemas de verificação de assinaturas pode ser realizada em função dos erros cometidos quanto à classificação. A Figura 3.8, apresenta duas distribuições normais para as classes genuína e falsificações. Desse modo, verifica-se uma intersecção entre essas curvas. A parte interseccionada representa os erros cometidos pelo sistema.

A partir do problema envolvendo duas classes, pode-se extrair duas métricas para avaliação de desempenho. Taxa de Falsa Rejeição (FRR) ou erro tipo I, nesse caso uma assinatura genuína é rejeitada pelo sistema e erroneamente classificada como falsa.

$$FRR = \frac{\textit{Número de Assinaturas Genuínas Rejeitadas}}{\textit{Número de Assinaturas Genuínas Submetidas}}$$

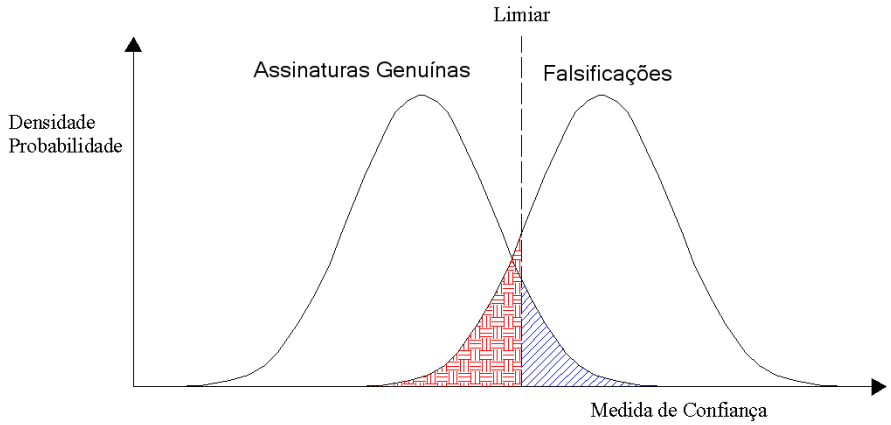


Figura 3.8: Intersecção existente entre assinaturas genuínas e falsificações.

Taxa de Falsa Aceitação (FAR) ou erro tipo II, ou seja, uma falsificação é incorretamente aceita pelo sistema e classificada como uma assinatura genuína.

$$FAR = \frac{\textit{Número de Falsificações Aceitas}}{\textit{Número de Falsificações Submetidas}}$$

Alguns trabalhos como [Sargur N. Srihari and Shah, 2007] [Coetzer, 2005] apresentam o Erro Médio (AER), o qual representa a média entre FAR e FRR.

$$AER = \frac{FRR + FAR}{2}$$

Contudo, um importante ponto a avaliarmos em sistemas de verificação de assinaturas é a necessidade e a importância de qual tipo de erro se deseja otimizar. Dessa forma pode-se ter sistemas no qual o preço pago por rejeitar uma assinatura verdadeira não seja tão alto, no qual a prioridade do sistema possa ser em não deixar assinaturas falsas serem classificadas como genuínas.

3.6 Curvas ROC

As curvas ROC (*Receiver Operating Characteristic*) têm sua origem fundamentada na teoria de sinais, sendo utilizadas há tempos na área médica. Após algum tempo pesquisadores da área de aprendizagem de máquina e reconhecimento de padrões passaram a utilizá-las no campo da informática apresentando, assim, um novo método para avaliação de algoritmos e classificadores, principalmente com classificação envolvendo duas classes. Apresenta-se a seguir uma contextualização sobre curvas ROC embasado no trabalho de Fawcett [Fawcett, 2006].

Ao observar um problema envolvendo duas classes (como o apresentado nesse trabalho), tem-se então, uma classe de interesse, a qual chamaremos de Positiva, e a outra de Negativa. Assim, para cada classe positiva/negativa, pode existir as duas classes (positiva e negativa). A Figura 3.9 demonstra estas quatro possíveis situações.

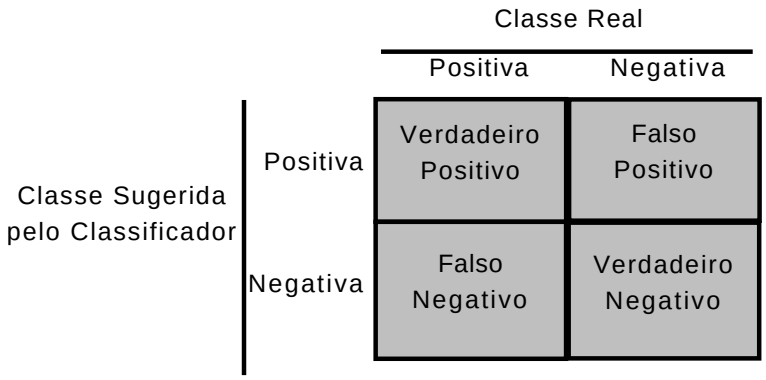


Figura 3.9: Quatro situações possíveis em um classificador a partir de duas classes.

- No caso de uma amostra Positiva ser classificada como Positiva, contabiliza-se então uma amostra Verdadeira Positiva. *(True Positives - **TP**)*.
- Uma amostra Positiva sendo classificada como Negativa, será contada como Falso Negativo. *(False Negatives - **FN**)*.
- Uma amostra Negativa sendo classificada como Negativa, é contada como Verdadeiro Negativo. *(True Negatives - **TN**)*.
- Por fim, para uma amostra Negativa sendo classificada como Positiva, é contada como Falso Positivo. *(False Positives - **FP**)*.

A Tabela 3.1 apresenta algumas métricas úteis derivadas da Figura 3.9.

Tabela 3.1: Métricas utilizadas em problemas com duas classes.

Nome	Métrica
Taxa de verdadeiros positivos (*Recall*)	$\frac{TP}{P}$
Taxa de falsos positivos	$\frac{FP}{P}$
Precisão	$\frac{TP}{TP+FP}$
Exatidão	$\frac{TP+TN}{P+N}$

Um gráfico de curvas ROC apresenta simplesmente a relação entre a taxa de TP (Verdadeiros Positivos) $\times$ FP (Falsos Positivos). Desse modo tem-se um gráfico bi-dimensional *(x,y)*, onde o eixo x apresenta a taxa de Falso Positivo e o eixo y a taxa de Verdadeiro Positivo.

Existem duas maneiras para a associação de amostras a classe, seja, de maneira discreta ou através de um *score*. De maneira discreta (menos utilizada), cada classificador apresenta um par (FP, TP) correspondente a um único ponto no gráfico ROC, como apresenta a Figura 3.10. Todavia, como o mesmo não emite uma nota durante a classificação, a função de decisão passa a ser discreta. Sendo assim, não existe maneira de variar a função de decisão para a análise de outras taxas para FP e TP. Deste modo o desempenho de cada classificador passa a ser representado por pontos no gráfico, conforme ilustrado na Figura 3.10.

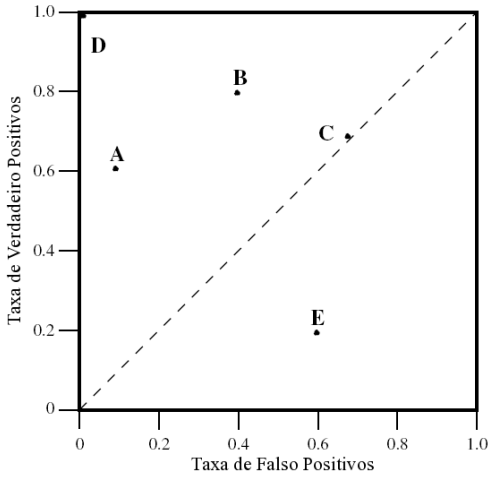


Figura 3.10: Gráfico ROC apresentando cinco classificadores discretos. Adaptado de [Fawcett, 2006].

Analisando a Figura 3.10, pode-se fazer algumas importantes observações. Por exemplo, os pontos *(0,0)* e *(1,1)*, têm por propriedade nunca apresentarem, respectivamente, classificação Positiva e Negativa. Já o ponto D (0,1) representa um classificação perfeita.

Diferente da maneira discreta, a função de decisão para classificadores baseados em *scores* (notas), ao invés de apresentarem um ponto no espaço ROC, apresentam um limiar. Dessa forma, entende-se que todas as notas que estiverem acima do limiar *score*, serão classificadas como Positivas, caso contrário como Negativas. Conceitualmente, os limiares existentes são infinitos, encontrando-se num espaço de $[-\infty$ até $+\infty]$. Assim o gráfico ROC será apresentado como uma curva. A Figura 3.11 demonstra um típico gráfico de curvas ROC.

Após construído os pontos ou curva no gráfico ROC, uma importante característica é avaliar o desempenho do classificador, o que é bastante simples, pois, quanto mais próxima do canto superior esquerdo a curva ou o ponto se encontrar, melhor é o desempenho do classificador. Desse modo, na Figura 3.10 onde foram apresentados alguns pontos, o ponto D (0,1), é o ponto ótimo, pois toca o canto superior esquerdo. Demonstra-se para o pior caso uma curva diagonal no gráfico ROC, ou seja, o desempenho é o mesmo que distribuir rótulos aleatórios, como um jogo de “cara ou coroa”.

Um método bastante utilizado na avaliação de desempenho é a fixação de pontos para FP. Por exemplo, avaliando a Figura 3.11 ao fixarmos o ponto FP = 0,1 (10%) teremos para TP uma taxa de 0,86. Contudo a fixação de um dado ponto é inerente ao problema em questão.

3.6.1 Área Abaixo da Curva ROC (AUC)

AUC, do inglês *Area Under Curve*, é uma métrica comumente citada e utilizada para comparar performance de curvas ROC. A área abaixo da curva é um dos índices mais utilizados para sumarizar a “qualidade” da curva. A AUC é uma descrição unidimensional da performance do classificador.

A princípio considerava-se a área total do gráfico ROC, observando que tais valores variavam de 0 a 1. Todavia, como descrito anteriormente, em que o pior caso seria uma linha diagonal entre (0,0) e (1,1) representando assim o uso de dados aleatórios, passou-se então a

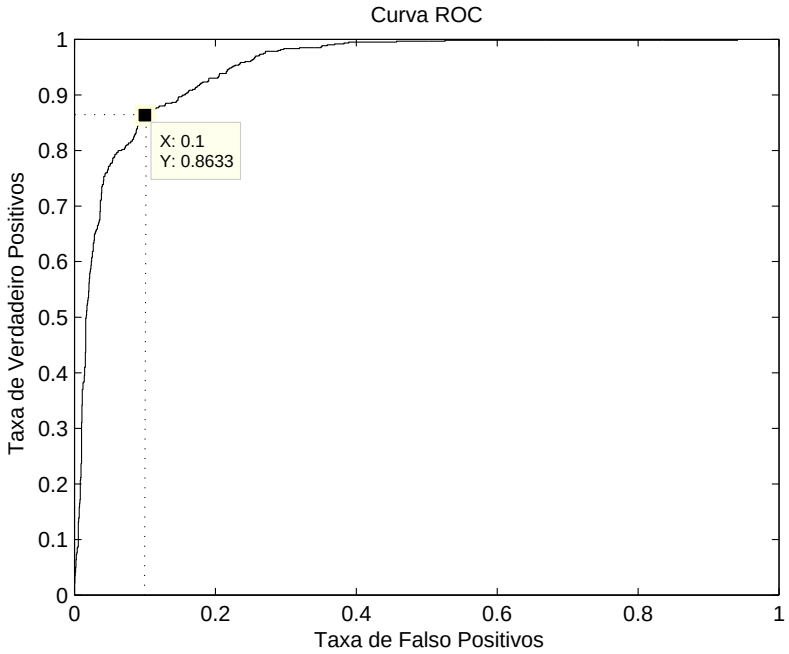


Figura 3.11: Típica curva ROC

não considerar valores inferiores a 0,5. Sendo assim, observa-se que para uma curva ótima seu AUC é 1, e para o pior caso (aleatório) o AUC é igual a 0,5, tocando a diagonal principal do gráfico.

Observando as áreas abaixo da curva dos classificadores A e B, representada através de um exemplo ideal pela Figura 3.12, percebe-se claramente que a área ocupada pelo classificador B é maior que a do classificador A, sendo assim, o desempenho médio de B é melhor que o de A. É possível também observar que mesmo um classificador possuindo um elevado AUC possa apresentar piores taxas em uma região específica do ROC que um classificador com baixa AUC. A Figura 3.12 demonstra esta hipótese, na qual para taxas de falsos positivos maior que 0,6 *(FP > 0,6)* o classificador A tem uma ligeira vantagem sobre B. Contudo, o exemplo ilustrado trata-se de um caso hipotético e ideal, podendo ser bastante diferente de exemplos práticos e reais.

3.7 Esquemas de Fusão

Apresenta-se a seguir uma síntese sobre esquemas de fusão de classificadores. Essa analise está embasada no reconhecido trabalho de Josef Kittler *et al.* [Kittler et al., 1998].

O motivo que leva a métodos de combinar classificadores são a eficiência e a precisão. A idéia de se combinar classificadores é não confiar em um único sistema de tomada de decisão, ao invés disso, todos os conjuntos ou subconjuntos são utilizados para a tomada de decisões, combinando suas decisões individuais a fim de obter uma decisão consensual.

Um aspecto interessante é a forma como podemos combinar classificadores. Caso possua-se apenas os rótulos disponíveis, a regra do voto majoritário pode ser utilizada. Tendo *scores* ou estimativas de probabilidades como saídas, alguma combinação linear é sugerida [Kittler et al., 1998].

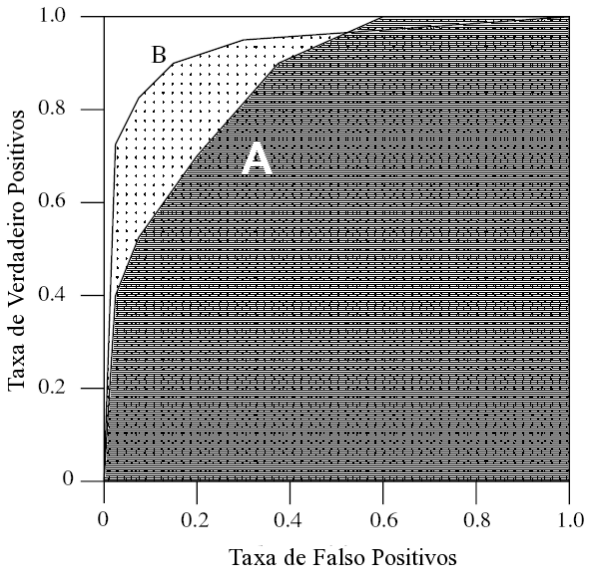


Figura 3.12: Gráfico ROC. Área abaixo da curva (situação hipotética). Adaptado de [Fawcett, 2006].

Os esquemas de fusão a serem tratados neste trabalho são métodos que independem dos dados, ou seja, não são influenciados por dados do treinamento, dessa forma os esquemas de fusão abordados aqui serão esquemas de agregação simples, sendo estes: Voto Majoritário, Soma, Produto, Média, Mediana, Máximo e Mínimo.

De acordo com o trabalho de Kittler *et al.* [Kittler et al., 1998], considerando um problema de reconhecimento de padrões em que assumimos o padrão *Z* para uma das *m* possíveis classes ($\omega_1, ..., \omega_m$). Supondo que existem *R* classificadores, no qual cada um representa uma classe por um vetor de atributos distintos e assumindo que o vetor usado pelo *i-ésimo* classificador é x_i. Na dimensão do espaço cada classe ω_k é modelada por uma função densidade de probabilidade $p(x_i|\omega_k)$ sendo sua probabilidade *a priori* de ocorrência denotada por $P(\omega_k)$.

Conforme a teoria Bayesiana, dada a dimensão $x_i, i = 1, ..., R$, o padrão *Z* deve ser atribuído a classe ω_j, na qual oferece a probabilidade *a posteriori* cuja a interpretação é máxima, ou seja:

$$\begin{aligned} &atribuir\ Z \rightarrow w_j\ se \\ P(w_j|x_1, ..., x_R) &= max_k P(w_k|x_1, ..., x_R) \end{aligned} \tag{3.19}$$

A regra de decisão de Bayes 3.19, estabelece que para utilizar toda a informação existente para se chegar a uma decisão correta, é essencial calcular as probabilidades de várias hipóteses, considerando simultaneamente todas as medidas.

Para computar a probabilidade *a posteriori* dependemos do conhecimento de medidas estatísticas de alta ordem, descritas em termos de funções de densidade de probabilidade conjunta $p(x_i, ..., x_R|\omega_k)$, que seria difícil para inferir. Tentando simplificar a regra 3.19 e exprimila em termos de apoio à decisão dos classificadores individuais, em que cada um explora somente as informações dadas pelo seu vetor de característica x_i. Sendo assim, consegue-se construir uma regra de decisão computacional mais eficiente, através de regras de combinação que são comumente utilizados na prática.

Dessa forma, ao reescrever a probabilidade *a posteriori* $p(\omega_k|x_1, ..., x_R)$, utilizando o teorema de Bayes, teremos:

$$P(w_k|x_1, ..., x_R) = \frac{p(x_1, ..., x_R|w_k)P(w_k)}{p(x_1, ..., x_R)} \tag{3.20}$$

no qual $p(x_1, ..., x_R)$ é uma medida absoluta da densidade de probabilidade conjunta. É apresentado então uma medida de distribuição condicional

$$P(x_1, ..., x_R) = \sum_{j=1}^{m} p(x_1, ..., x_R|w_j)P(w_j) \tag{3.21}$$

para adiante, considera-se a Equação 3.20 como base.

3.7.1 Regra do Produto

Tem-se que $p(x_1, ..., x_r|\omega_k)$ representa o conjunto de distribuições de probabilidade das medidas extraídas pelos classificadores. Define-se a regra do produto a partir de (3.22)

$$p(x_1, ..., x_R|w_k) = \prod_{i=1}^{R} p(x_i|w_k) \tag{3.22}$$

no qual $p(x_i|\omega_k)$ é o modelo da *i*-ésima representação. Deste modo, substituindo (3.22) e (3.21) em (3.20) encontra-se:

$$P(w_k|x_1, ..., x_R) = \frac{P(w_k)\prod_{i=1}^{R} p(x_i|w_k)}{\sum_{j=1}^{m} P(w_j)\prod_{i=1}^{R} p(x_i|w_j)} \tag{3.23}$$

e aplicando (3.24) em (3.19), obtem-se a regra de decisão do produto

$$atribuir\ Z \rightarrow w_j\ se$$
$$P(w_j)\prod_{i=1}^{R} P(x_i|w_j) = max_{k=1}^{m} P(w_k)\prod_{i=1}^{R} p(x_i|w_k) \tag{3.24}$$

ou em termos de probabilidade *a posteriori* fornecidas pelos respectivos classificadores

$$atribuir\ Z \rightarrow w_j\ se$$
$$p^{-(R-1)}(w_j)\prod_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} P^{-(R-1)}(w_k)\prod_{i=1}^{R} P(w_k|x_i) \tag{3.25}$$

A regra de decisão (3.25) quantifica a probabilidade de uma hipótese ser combinana com a probabilidade *a posteriori* geradas por classificadores indivíduais através da regra do produto. Em suma, concluí-se que esta é uma regra eficientemente severa.

3.7.2 Regra da Soma

Para a regra da soma, considera-se a regra do produto (3.25) em maiores detalhes. Em alguns casos assume-se que a probabilidade *a posteriori* calculada pelo respectivo classificador não diferenciará drasticamente da probabilidade *a priori*. Essa hipótese pode ser satisfeita quanto à disposição da informação ser muito ambígua devido ao alto nível de ruído. Nessa situação pode-se assumir que a probabilidade *a posteriori* pode ser expressa em

$$P(w_k|x_i) = P(w_k)(1 + \delta_{ki}) \tag{3.26}$$

no qual δ_{ki} safisfaça $\delta_{ki} << 1$. Assim, substituindo (3.26) nas probabilidades *a posteriori* em (3.25), encontra-se:

$$p^{-(R-1)}(w_k) \prod_{i=1}^{R} P(w_k|x_i) = P(w_k) \prod_{i=1}^{R} (1 + \delta_{ki}) \tag{3.27}$$

Ao expandir o produto e negando os termos de segunda ordem, podemos aproximar o lado direito da equação (3.27) assim:

$$P(w_k) \prod_{i=1}^{R} (1 + \delta_{ki}) = P(w_k) + P(w_k) \sum_{i=1}^{R} \delta_{ki} \tag{3.28}$$

Por fim, substituindo (3.28) e (3.26) em (3.25) obtém-se então a regra de decisão da soma:

$$(1 - R)P(w_j) + \sum_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} \left[(1 - R)P(w_k) + \sum_{i=1}^{R} P(w_k|x_i) \right] \quad \begin{array}{r} atribuir\ Z \rightarrow w_j\ se \end{array} \tag{3.29}$$

Kittler [Kittler et al., 1998] faz um breve reflexão sobre as regras da soma e do produto, em que talvez o ponto mais importante é o fato de todas as regras de decisões a serem derivadas destas são largamente usadas na prática.

3.7.3 Esquemas de Combinação de Classificadores

Através dos esquemas de combinação apresentados por Kittler, a partir das regras de decisão do produto (3.25) e da soma (3.29) outras regras de decisão podem ser desenvolvidas, observando que

$$\prod_{i=1}^{R} P(w_k|x_i) \leq min_{i=1}^{R} P(w_k|x_i) \leq \frac{1}{R} \sum_{i=1}^{R} P(w_k|x_i) \leq max_{i=1}^{R} P(w_k|x_i) \tag{3.30}$$

A relação (3.30) sugere que as regras da soma e do produto possam ser aproximadas para seus limites (máximos e mínimos). O enrijecimento das probabilidades *a posteriori* $P(K_i|x_i)$ para produzir funções com valores binários.

$$\Delta_{ki} = \begin{cases} 1 & se\ P(w_k|x_i) = max_{j=1}^{m} P(w_j|x_i) \\ 0 & caso\ contrário \end{cases} \tag{3.31}$$

Regra do Máximo

A partir de (3.29) é possível aproximar a regra da soma pelas máximas probabilidades *a posteriori*

$$atribuir\ Z \rightarrow w_j\ se$$
$$(1-R)P(w_j) + Rmax_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} \left[(1-R)P(w_k) + Rmax_{i=1}^{R} P(w_k|x_i)\right] \tag{3.32}$$

onde assumindo probabilidades *a priori* iguais, obtém-se a regra do máximo 3.33:

$$atribuir\ Z \rightarrow w_j\ se$$
$$max_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} max_{i=1}^{R} P(w_k|x_i) \tag{3.33}$$

Regra do Mínimo

A partir de (3.25) limitando o produto das probabilidades *a posteriori* pelo seu máximo

$$atribuir\ Z \rightarrow w_j\ se$$
$$p^{-(R-1)}(w_j) min_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} p^{-(R-1)}(w_k) min_{i=1}^{R} P(w_k|x_i) \tag{3.34}$$

onde assumindo probabilidades *a priori* iguais, obtém-se a regra do mínimo

$$atribuir\ Z \rightarrow w_j\ se$$
$$min_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} min_{i=1}^{R} P(w_k|x_i) \tag{3.35}$$

Regra da Mediana

Supondo que a regra da soma (3.29) com o mesmo conhecimento *a priori* para as classes, pode ser computada e visualizada através da média das probabilidades *a posteriori* para cada classe, durante a saída dos classificadores, tem-se,

$$atribuir\ Z \rightarrow w_j\ se$$
$$\frac{1}{R}\sum_{i=1}^{R} P(w_j|x_i) = max_{k=1}^{m} \frac{1}{R} P(w_k|x_i) \tag{3.36}$$

Dessa forma, a regra da mediana atribui um padrão à classe cuja probabilidade *a posteriori* seja máxima. Entretanto, se um dos classificadores de saída adotar uma probabilidade *a posteriori* com um desvio muito grande das demais classes, isso afetará a média podendo conduzir a uma decisão incorreta. Sabendo disso, o mais adequado é basear a decisão de combinação na mediana das probabilidade *a posteriori*, levando para a seguinte regra:

$$\begin{aligned} &atribuir\ Z \rightarrow w_j\ se \\ med_{i=1}^{R} P(w_j|x_i) &= max_{k=1}^{m} med_{i=1}^{R} P(w_k|x_i) \end{aligned} \tag{3.37}$$

Regra do Voto Majoritário

A partir de (3.29) assumindo a probabilidade, *a priori*, e o enrijecimento das probabilidades de acordo com (3.31), encontra-se

$$\begin{aligned} &atribuir\ Z \rightarrow w_j\ se \\ \sum_{i=1}^{R} \Delta_{ji} &= max_{k=1}^{m} \sum_{i=1}^{R} \Delta_{ki} \end{aligned} \tag{3.38}$$

Com relação à soma do lado direito da equação (3.38) para cada w_k tem-se a contagem dos votos recebidos para dada hipótese dos classificadores individuais. Assim, a classe com o maior número de votos é selecionada pelo (consenso ou decisão) da maioria.

3.8 Algoritmos Genéticos

Computação Evolutiva, é o nome dado aos estudos embasados na teoria Darwiniana aplicada à informática. O tema já vem sendo estudado desde a década de 50. *Adaptation in Natural and Artificial Systems* [Holland, 1992], é uma obra clássica sobre o assunto, escrita por John H. Holland, considerado o pai dos Algoritmos Genéticos (AGs). David Goldberg e Lawrence Davis, são outros pesquisadores que contribuíram e contribuem muito com o tema.

Algoritmos Genéticos fazem parte da família da Computação Evolutiva, esses são modelos computacionais inspirados no processo evolutivo biológico. Basicamente, a maior parte das variações de algoritmos genéticos funcionam como mecanismo de busca e otimização. Isso se deve a sua versatilidade para resolução de problemas complexos.

AGs são muito eficientes para busca de soluções ótimas, ou sub-ótimas em uma grande variedade de problemas, pois não impõem muitas das limitações encontradas nos métodos de busca tradicionais. Isso acontece devido aos métodos tradicionais utilizarem regras de transição probabilísticas e não determinísticas, implicando assim, na redução de chances do operador ficar preso em ótimo local, convergindo para um ótimo global, conforme ilustra a Figura 3.13.

Tem-se então que AGs utilizam um sistema de busca paralela, estruturada e randômica, em que buscam-se pontos de alta aptidão. Pontos de alta aptidão, do inglês *fitness*, são pontos na função a qual se deseja minimizar ou maximizar.

A inicialização da população com valores randômicos é o primeiro passo no ciclo de um AG. A técnica de inicializar um indivíduo da população com uma solução conhecida, sabendo que a solução desse se encontra perto de uma solução ótima, é comumente utilizada.

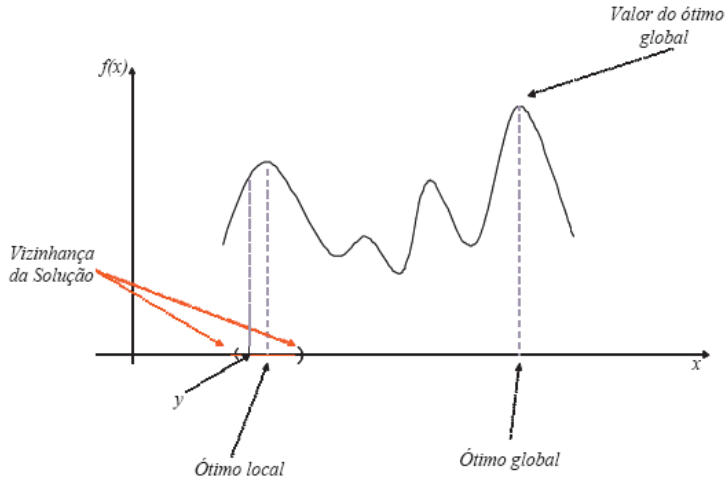


Figura 3.13: Exemplo típico de ótimo local e ótimo global.

Posteriormente, é necessário calcular a aptidão de cada indivíduo da população. O valor da aptidão é um parâmetro importante para o algoritmo, pois na seleção de indivíduos usados para a reprodução, que dará origem a uma nova geração, é geralmente baseada nesse parâmetro. Assim, quanto maior a aptidão de um indivíduo, maior são suas chances de ser selecionado para ser um reprodutor.

Em seguida, a população é submetida a operações genéticas do tipo cruzamento e mutação. Finaliza-se quando um dos procedimentos de parada é alcançado, podendo ser o número de iterações, o tempo ou quando um indivíduo alcança uma aptidão pré-estabelecida.

Após estes quatro passos: (1) Inicialização; (2) Cálculo da aptidão; (3)Geração de uma nova população e (4) Operações genéticas, se essas condições não forem satisfeitas, temos o retorno ao cálculo da aptidão, passando por todas as etapas, até que uma condição de parada pré-determinada seja encontrada, conforme demonstra o diagrama de estados da Figura 3.14.

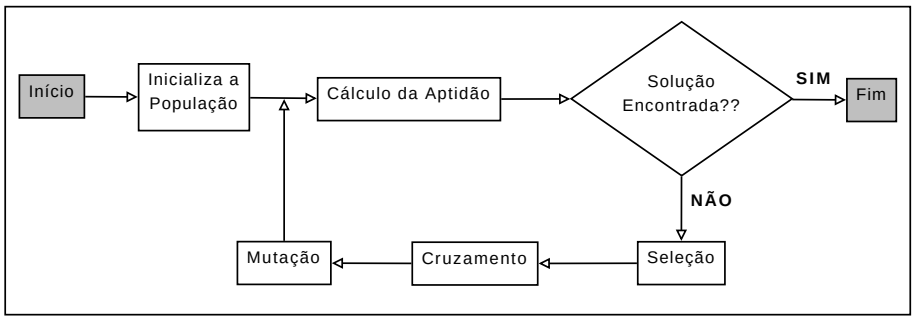


Figura 3.14: Ciclo do Algoritmo Genético.

O Algoritmo 1 apresenta o pseudo-código de um clássico algoritmo genético.

Algoritmo 1 Algoritmo Genético Clássico

1: $t \leftarrow 0$
2: Inicializar População *(t)*
3: **while** condição de término não for satisfeita **do**
4: $t \leftarrow t + 1$
5: Seleciona a População' *(t)* a partir da População *(t -1)*
6: Cruzamento População' *(t)* para gerar a População"*(t)*
7: Mutação da População" *(t)* para gerar a População"'*(t)*
8: Avaliação da População"' *(t)*
9: **end while**

3.8.1 Componentes de um AG Clássico

Basicamente no funcionamento de um AG clássico temos três tipos de operaçõs: seleção, cruzamento e mutação. Para um maior entendimento das operações que compões um AG descreve-se com maiores detalhes algumas operações presentes em um algoritmo genético, são elas:

Aptidão (*Fitness*)

A aptidão de um indivíduo é o parâmetro que determina quão boa é a solução por esse apresentada para o problema em questão. É a função de aptidão que determina se o indivíduo, com uma certa taxa de aptidão pode ou não fazer parte de uma população futura. Para escolha da função aptidão temos de estar intimamente ligados ao domínio do problema. Nos testes realizados nesta pesquisa, utiliza-se três parâmetros como aptidão, a fim de avaliar o impacto de cada um.

Tamanho da População

Com relação ao tamanho da população vemos que ela está fortemente ligada ao problema (espaço de busca). Normalmente empregam-se populações variando de 20 a 200 indivíduos, porém isso é relativo e alterado de acordo com a complexidade do problema a ser tratado. A principal característica do tamanho da população é a busca realizada no problema, dessa forma quanto maior o tamanho da população, mais completa será a busca realizada pelo algoritmo e mais cara computacionalmente.

Métodos de Seleção

Diversos são os métodos de seleção existentes e propostos atualmente. O método de seleção clássico de um AG é o da roleta. Nesse método tem-se uma relação direta com a aptidão do indíviduo, ou seja, quanto maior a aptidão do indivíduo, maior são as chances de ele fazer parte de uma nova população. O que ocorre é que a chance de um indivíduo passar para a próxima geração é proporcional a sua aptidão medida em relação à soma das aptidão de todos os indivíduos. A chance dos indivíduos que tem as maiores aptidões passarem para uma próxima geração são muito maiores que a de indivíduos com baixa aptidão. No entanto, com o método da roleta, pode ser que o melhor indivíduo de uma população seja descartado, ou seja, esse indivíduo pode não passar a fazer parte da nova população. Alguns outros métodos

existentes são: Estocástico, Uniforme e *Ranking*. Neste trabalho, utiliza-se como mecanismo de seleção o Método de *Ranking* citado por Back [Bäck, 1996].

Mutação

O operador genético de mutação consegue introduzir uma maior variabilidade dentro de uma população. Consiste basicamente na alteração de um ou mais genes de forma aleatória dentro do cromossomo. A característica do operador de mutação é que se crie uma diversidade extra na população, porém sem acarretar danos ao progresso construído com a busca. No exemplo a seguir temos uma cadeia de 0's e 1's de comprimento igual a 8. Se aplicarmos o operador de mutação no quinto elemento dessa cadeia, teríamos o seguinte indivíduo gerado:

indivíduo Inicial: [0 1 1 1 ***1*** *1 1 0]*
indivíduo Gerado: [0 1 1 1 ***0*** *1 1 0]*

Um dos parâmetros do AG é a probabilidade de mutação, o qual é a probabilidade que um gene mute ou se altere. Os valores utilizados para a taxa de mutação geralmente são pequenos (1% a 10%). Os métodos de mutação mais conhecidos são a mutação uniforme e mutação Gaussiana.

Cruzamento

A idéia base para o operador de cruzamento é a criação de novos indivíduos a partir da combinação de dois ou mais indivíduos pais. Com isso temos a troca de informações entre diferentes soluções candidatas. Na proposta de cruzamento de um ponto, observamos que a partir da troca de características pertencentes a dois indivíduos pais a formação de dois indivíduos filhos, cujos segmentos de características dos indivíduos filhos são referentes aos indivíduos pais, contudo, nesse caso entende-se que um único ponto de corte é gerado aleatoriamente. O exemplo hipotético a seguir com uma cadeia de 8 bits de 0's e l's, vejamos:

pai_1 = [1 1 1 0 1 1 1 1]
pai_2 = [0 0 0 1 1 0 0 0]

Aplicando o operador de cruzamento a partir da quarta posição, teríamos então o cruzamento entre os dois indivíduos pais, surgindo assim os dois indivíduos filhos desse cruzamento:

pai_1 = [1 1 1 0 **1 1 1 1**]
pai_2 = [0 0 0 0 **1 0 0 0**]

indivíduos filhos gerados:

$filho_1$ = [1 1 1 0 **1 0 0 0**]
$filho_2$ = [0 0 0 0 **1 1 1 1**]

Valores dos Parâmetros

Muitos dos parâmetros utilizados nos experimentos com AG's foram determinados por tentativa e erro ou com base em outros trabalhos correlatos, como: tamanho da população, taxa de mutação, número de gerações, taxa de cruzamento, entre outros.

3.9 Agrupamento de Classificadores (*Ensembles*)

Segundo Dietterich [Dietterich, 2000], agrupamento de classificadores é:

> *"Um conjunto de classificadores cujas decisões individuais são combinadas (através de algum método), a fim de classificar novos exemplos"*.

A combinação de classificadores pode apresentar melhores taxas que a taxa de um classificador individual. Entretanto, para agrupamentos de classificadores apresentarem resultados mais precisos que os de classificadores isolados, necessita-se que os classificadores sejam distintos. Podemos classificar como sendo classificadores distintos aqueles que, dado um exemplo, cada classificador cometa erros diferentes. Essa diversidade faz com que os agrupamentos apresentem uma maior precisão se comparados aos classificadores isolados [Hansen and Salamon, 1990].

Dietterich [Dietterich, 2000] apresenta um simples e interessante exemplo de como a combinação de classificadores pode melhorar as taxas. No exemplo, tem-se um agrupamento formado por 3 classificadores distintos (C_1, C_2 e C_3). Observa-se também um novo exemplo *x* a ser classificado. A Figura 3.15, demonstra tal idéia.

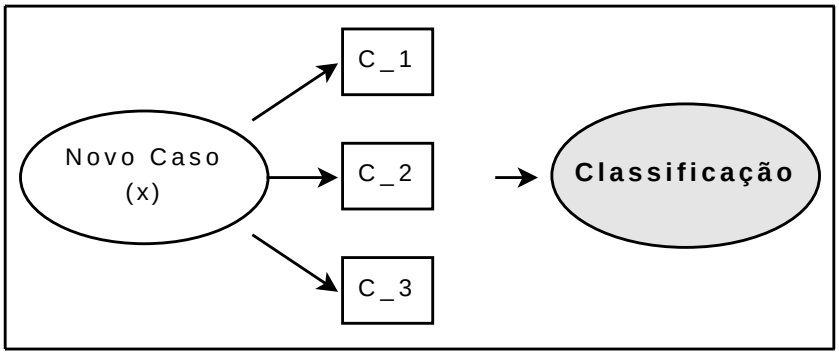


Figura 3.15: Exemplo de um *ensemble* formado por 3 classificadores distintos

Se os três classificadores forem idênticos, então quando, $C_1(x)$ for incorreto, logo $C_2(x)$ e $C_3(x)$ também serão incorretos. No entanto, se os classificadores forem distintos ou não correlatos, quando $C_1(x)$ for incorreto, C_2 e C_3 podem ser corretos. Deste modo, poderíamos utilizar o esquema de voto majoritário, e classificar o exemplo *x* corretamente. O autor ainda faz um estudo acerca das probabilidades de um agrupamento composto por *L* classificadores $C_1, ..., C_L$ com taxas de erros menores e maiores que 50%.

Bernardini [Bernardini, 2006] descreve e simplifica algumas idéias de Dietterich. Consideremos sempre que os classificadores a serem citados são não correlacionados. Descreveremos também que a probabilidade *e* é a taxa de erro proporcionada por cada classificador. Tem-se então, *p* como sendo uma classificação de *x* correta e (1 - *e*) como sendo uma classificação incorreta. Pode-se descrever, então, que em um agrupamento de *L* classificadores, uma probabilidade *l* de sucessos, pode ser dada através de:

$$E(Z = l) = \left(\frac{L}{l}\right) e^l (1 - e)^{L-l} \tag{3.39}$$

Se no caso hipotético, as taxas de erro dos L classificadores forem sempre menores que 50% ($e<\frac{1}{2}$), sendo ainda que tais erros sejam independentes, pode-se então demonstrar que a chance do método do voto majoritário classificá-lo erroneamente é dada através de:

$$E(Z > \frac{L}{2}) = 1 - \sum_{l=1}^{\frac{L}{2}} \left(\frac{L}{l}\right) e^l (1-e)^{L-l} \tag{3.40}$$

Demonstra que mais da metade dos L correspondem a classificadores errados. A seguir, um exemplo hipotético ideal será demonstrado, justificando que o uso de um esquema de combinação pode contribuir quando existem taxas de erro menores que 50% e piorar quando essas taxas ultrapassam os 50%.

Dietterich [Dietterich, 2000] demostra em seu trabalho que dado um determinado exemplo hipotético com 21 hipóteses independentes cujas taxas de erros são de 0,3, 0,4 e 0,45, verifica-se que quanto maior a taxa de erro pertencente às hipóteses, maior será a taxa de erro do agrupamento, contudo a taxa de erro dada pelo agrupamento é muito menor que a taxa apresentada pelas hipóteses independentes.

Bernardini [Bernardini, 2006] embasada no exemplo de Dietterich demonstra através de um gráfico o ganho de desempenho quando combinados os classificadores. Nesse caso, através das 21 hipóteses simuladas, com taxas de erro de 0,3, demonstra o gráfico que para 11 ou mais hipóteses simultaneamente incorretas a taxa de erro no gráfico cai para 0,026, sendo muito menor que a taxa proporcionada por hipóteses individuais, veja a Figura 3.16.

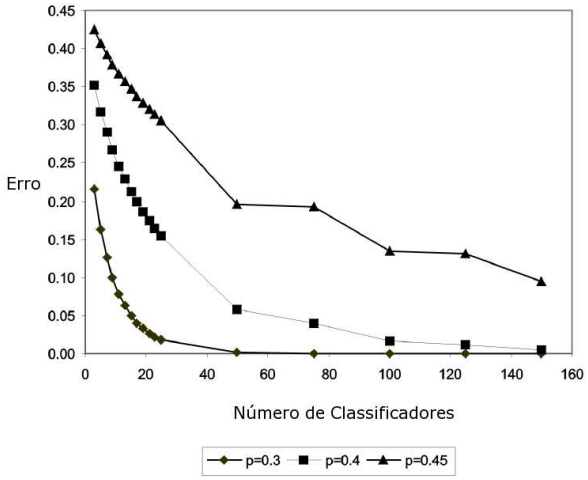


Figura 3.16: Desempenho com o uso de *ensembles* onde as taxas de erros dos classificadores eram menores que 0.5

Observando a Figura 3.16 fica claro que: (1) Quanto maior o número de hipóteses independentes existentes, menor é a taxa de erro do agrupamento; (2) Quanto menor a taxa de erro das hipóteses, menor é a taxa de erro do agrupamento.

Capítulo 4

Metodologia Proposta

Neste capítulo apresenta-se a metodologia utilizada no desenvolvimento deste trabalho. A Figura 4.1 apresenta uma idéia geral dos procedimentos a serem realizados e a seguir cada uma dessas etapas será descrita em detalhes.

4.1 Definição do Problema

Criar agrupamento de classificadores baseados em dissimilaridade para reduzir taxas de erro tipo I e II em problemas de verificação de assinaturas *off-line*. Existem disponíveis quatro conjuntos de características, os quais foram usados para treinar 64 classificadores SVM (referente a quatro características combinadas com 16 diferentes configurações de *grids*), sendo seus objetivos discriminar assinaturas genuínas de falsificações (Figura 4.2). Esses classificadores serão combinados com o auxílio de um algoritmo genético e diferentes funções de aptidão serão consideradas.

4.2 Definição da Base de Dados

Com relação à base de dados, essa se encontra disponível no Laboratório de Visão, Imagem e Robótica (VIR) do PPGIa - PUCPR. Um maior detalhamento quanto a aquisição e etapas de pré-processamento da base, juntamente com normas impostas pelo Banco Central a modelos de cheques pode ser consultada na tese de Justino [Justino, 2001].

4.2.1 Aquisição dos Dados

A base disponível no laboratório de Imagem, Visão e Robótica (VIR), possui 5200 imagens de assinaturas. Tais imagens foram recortadas de tamanhos fixos 3x10 cm (tamanho dedicado a assinatura em cheques bancários), com 256 níveis de cinza e densidade de 300 dpi, salvas em arquivos do tipo BMP.

Para chegar a este número de imagens de assinaturas, foram utilizados 100 autores. Cada autor cede 40 assinaturas originais. Dos 100 autores, os 60 primeiros cedem também 10 falsificações simples e 10 falsificações simuladas.

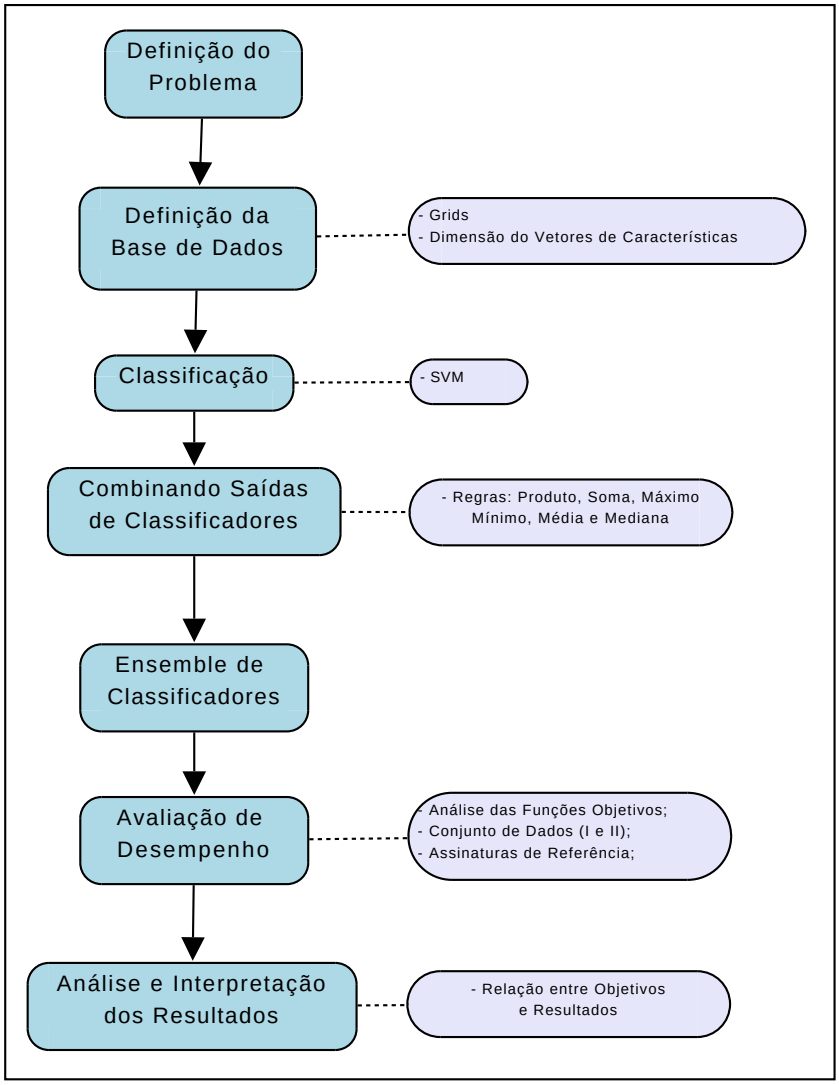


Figura 4.1: Metodologia proposta.

(a) (b) (c)

Figura 4.2: Exemplos de assinatura: (a) genuína, (b) falsificação simples, e (c) falsificação simulada.

Um dos entraves na criação de bases de assinaturas é a quantidade de assinaturas coletadas por cada autor. Santos [Santos, 2004], descreve a necessidade de confeccionar bases com coletas de assinaturas programadas em períodos distribuídos.

Na prática, o número de assinaturas coletadas por autor são em torno de cinco. Em nossos experimentos avaliaremos o impacto proporcionado por diferentes números de assinaturas no processo de treinamento. A base de dados foi dividida em 40, 20 e 40 autores para treinamento, validação e testes, respectivamente.

4.2.2 Segmentação

A parte de segmentação é de grande valia e interesse para este trabalho. Isso ocorre por utilizar-se diferentes tamanhos de grades na geração dos classificadores. Brito *et al.*[Britto et al., 2001] cita as abordagens contextuais e não contextuais como sendo as principais abordagens de segmentação para assinaturas.

- **Contextual:** Nessa abordagem a idéia é identificar e utilizar as letras que constituem o nome do autor do modelo da assinatura. Essa técnica fica melhor adaptável em sistemas de reconhecimento de caracteres, já que esses tem formas mais coerentes, diferente de uma rubrica por exemplo.

- **Não Contextual:** Essa abordagem faz uso de características relacionadas à forma dos traços das assinaturas e leva em consideração aspectos geométricos e estatísticos desses traços.

A abordagem não contextual se mostra mais adequada para sistemas de verificação de assinaturas devido à grande quantidade do rubricas, estilo bastante utilizado. Esse método de segmentação permite a utilização de diferentes técnicas.

Uma técnica bastante difundida em diferentes trabalhos [Justino, 2001], [Oliveira et al., 2007], [Sabourin and Genest, 1995], é a utilização de grades (*grids*) sobreposta à imagem da assinatura. A Figura 4.3 apresenta duas diferentes configurações de *grids* para a mesma assinatura.

Devido aos segmentos existentes nas imagens de assinaturas possuírem formas e comprimentos variados, a avaliação destes em um contexto local é muito mais simples que em um contexto geral [Santos, 2004]. Características como espaços em branco existentes nas células de *grids*, podem ser de suma importância, podendo identificar por exemplo espaços entre dois ou mais blocos de uma assinatura.

O método de segmentação através de *grids* foi utilizado nas quatro diferentes características extraídas, assim todos os recursos são computados para cada célula do *grid*. Consideramos 16 diferentes variações de *grids* (Horizontal x Vertical):

Figura 4.3: Dois diferentes exemplos de configurações de *grids* usados para extração de características.

Tabela 4.1: Variações para tamanhos de *grids*

Horizontal × Vertical			
4×5	5×5	8×5	10×5
4×10	5×10	8×10	10×10
4×20	5×20	8×20	10×20
4×25	5×25	8×25	10×25

Nesse caso, a dimensão do vetor de características é totalmente relativa a configuração do *grid*. Isso pois, são computadas características referentes à cada célula do *grid*, assim um *grid* que tenha uma configuração de 4 × 5 é composto por 20 células, e seu vetor de características será relativo às 20 células. Utilizando um *grid* que tenha uma configuração de 10 × 25, teremos então 250 células, sendo assim, o vetor referente a esta configuração será muito maior.

4.2.3 Dimensão dos Vetores de Características

Para esse trabalho considera-se uma abordagem baseada na dissimilaridade, na qual os classificadores são treinados a fim de discriminar assinaturas genuínas (positivos) de falsificações (negativos), ver Seção 3.1. Para gerar as amostras positivas, são computadas a dissimilaridade entre quatro vetores de amostras genuínas de cada autor, resultando em seis possíveis combinações. Considerando que esse trabalho contra com 40 autores para o processo de treinamento, obtém-se um total de 240 exemplos positivos. A Figura 4.4 exemplifica esse processo.

Para construção das amostras negativas, utiliza-se as duas primeiras amostras dos 36 primeiros autores do conjunto de treinamento. Calcula-se a diferença entre essas assinaturas e outras duas assinaturas de quatro diferentes autores (selecionados ao acaso). Dessa forma totalizamos 288 amostras negativas. A Figura 4.5 ilustra o processo quanto às amostras negativas.

As 528 amostras disponíveis (positivas + negativas), baseadas na dissimilaridade a partir de nosso conjunto de treinamento, são utilizadas para alimentar o classificador SVM. Para o treinamento foi utilizada a validação cruzada *k-fold* com k = 10 e *kernel* RBF. Outros *kernels* foram testados, contudo, o RBF apresentou melhores resultados.

Na abordagem de dissimilaridade adotada, as amostras positivas e negativas no conjunto de teste dependem diretamente do número de assinaturas de referência (Sk). Ao utilizar como referência cinco assinaturas genuínas de um determinado autor (não sendo estas integrantes do conjunto de treinamento), calcula-se a diferença existente entre elas e de determinado vetor de teste (Sq), resultando em cinco vetores de dissimilaridades de características.

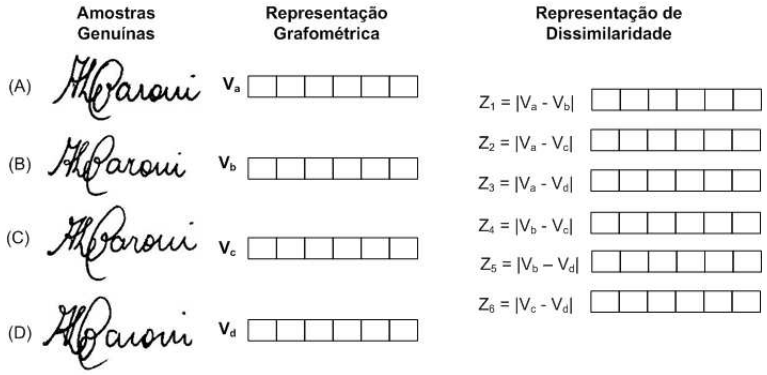


Figura 4.4: Dissimilaridades entre amostras genuínas do mesmo autor para gerar amostras positivas. A partir de quatro amostras genuínas, seis vetores de dissimilaridade são criados.

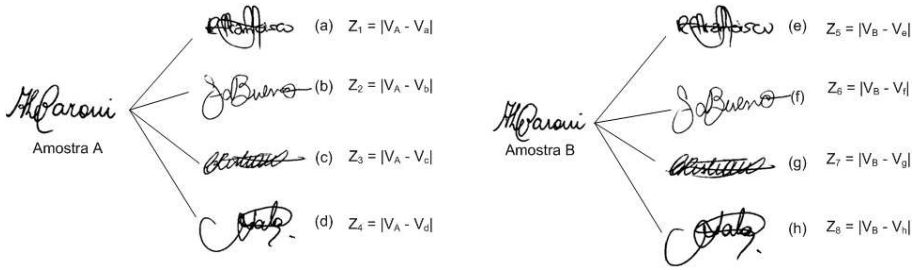


Figura 4.5: Dissimilaridade entre amostras genuínas de diferentes autores para gerar exemplos negativos.

Para o conjunto de testes, utilizou-se 40 autores com 40 amostras por autor, sendo 10 genuínas, 10 falsificações simples, 10 falsificações aleatórias (selecionadas aleatoriamente) e 10 falsificações simuladas. O número de assinaturas utilizadas como referência é objeto de estudo em nosso trabalho, sendo que para os testes utilizamos, 3, 5, 7, 9, 11, 13 e 15 assinaturas como referência.

4.3 Conjunto de Características

Descreve-se aqui as características das imagens de assinaturas utilizadas neste trabalho. Quatro características foram utilizadas nos testes, no entanto, trabalhos relatando sobre diversos tipos de características podem ser encontrados na literatura. As características utilizadas neste trabalho foram: Distribuição de *Pixels*, Curvatura, Densidade de *Pixels* e Inclinação. Apesar da extração de características propriamente dita não fazer parte do escopo do trabalho, uma breve descrição sobre as mesmas pode ajudar na compreensão dos resultados.

4.3.1 Distribuição de *Pixels*

Extended Shadow Code (ESC) é uma técnica inserida nos preceitos da abordagem local, utilizando-se de características estatísticas. Proposta por Sabourin e Genest [Sabourin and Genest, 1995], o método descreve uma técnica em que operações de projeções são realizadas sobre os *pixels* através de *grids* sobrepostos à imagem. Dessa forma, verifica-se

a ocorrência de cada *pixel* preto nas barras Horizontal, Vertical e Diagonal (HVD). Temos na projeção um prévio conjunto de *pixels* distribuídos ao longo das barras (HVD), e o que ocorre é a contagem destes *pixels*. Existe a necessidade de normalizar o número de *pixels* de cada projeção de acordo com o tamanho do *grid*. A partir daí, pode-se realizar um mapeamento da distribuição geométrica dos *pixels* na célula. Contudo, segundo Justino [Justino, 2001], os sensores em diagonal podem armazenar informações redundantes, optando-se assim por usar somente barras horizontais e verticais correspondentes às células dos *grids*, Figura 4.6.

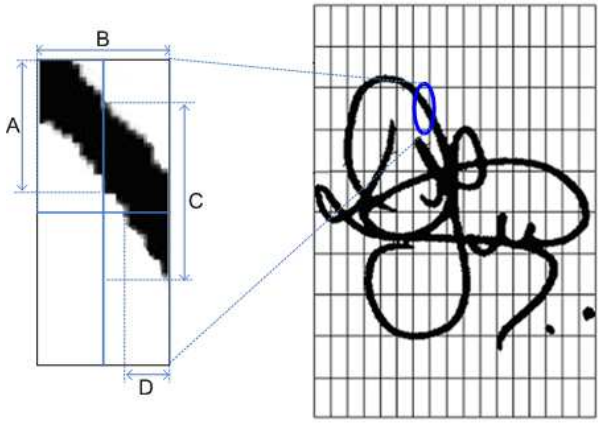


Figura 4.6: Exemplo do método de Distribuição de *Pixels*.

Todavia, o método proposto não incorpora características pseudodinâmicas presentes na assinatura.

4.3.2 Curvas de Bezier

Método proposto por Bertolini *et al.* [Bertolini et al., 2008], em que informações sobre a característica curvatura são obtidas através dos segmentos mais importantes presentes na assinatura. Para reproduzir tais segmentos foram usadas as curvas de Bezier [Sproull, 1979], as quais são definidas por quatro pontos: dois pontos de parada (origem e destino) e dois pontos de controle. Com o propósito de reduzir a complexidade do objetivo em questão, a imagem da assinatura é afinada, sendo delas extraídos contornos superiores e inferiores. Somente o traçado mais longo de cada célula será considerado para o processo de extração de característica. Para cada traçado, encontramos dois pontos finais e baseados neles, três pontos equidistantes (N) são definidos. A Figura 4.7 representa uma assinatura e seus respectivos contornos.

(a) (b)

Figura 4.7: (a) Assinatura genuína, e (b) Contornos da Assinatura.

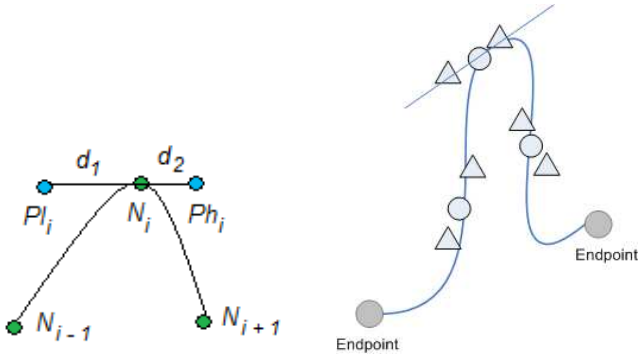
Para cada ponto $N_i, \{i = 1, 2, 3\}$ foi computada a tangente ($\tan N$) e os pontos de controle (Pl e Ph) através das Equações 4.1 e 4.2, respectivamente,

$$\tan N = \arctan \frac{y_{N-1} - y_{N+1}}{x_{N-1} - x_{N+1}} \tag{4.1}$$

$$\begin{cases} Pl_i(x) = & N_i(x) + \cos(\tan N_i) \times dist(N_i, N_{i-1}) \\ Pl_i(y) = & N_i(y) + \sin(\tan N_i) \times dist(N_i, N_{i-1}) \\ & \\ Ph_i(x) = & N_i(x) + \cos(\tan N_i) \times dist(N_i, N_{i-1}) \\ Ph_i(y) = & N_i(y) + \sin(\tan N_i) \times dist(N_i, N_{i-1}) \end{cases} \tag{4.2}$$

no qual $dist$ na Equação 4.2 representa a distância euclidiana. A Figura 4.8 apresenta um exemplo para as características computadas para N_i. A primeira é a tangente de N_i. Já o segundo e terceiro (d_1 e d_2) são as distâncias euclidianas de N_i para os dois pontos de controle, respectivamente. Na Figura 4.8a podemos notar a maior distância e a maior curvatura existente entre dois pontos do traçado.

Assim, são extraídas três características para cada ponto (tangente de N_i, d_1 e d_2), resultando em nove elementos para cada célula existente no *grid*. A Figura 4.8b ilustra um caso no qual os pontos são detectados em uma imagem de assinatura real (4.7b)



(a) (b)

Figura 4.8: (a) Exemplo de características extraídas do traçado e (b) exemplo de pontos detectados em um caso real, através da assinatura da Figura 4.7b.

4.3.3 Densidade de *Pixels*

Embasado nos trabalhos de Rigool e Kosmala, *appud* [Justino, 2001], a característica densidade de *pixels* contabiliza o número de *pixels* pretos (*Np*) existentes em cada célula do *grid* em uma imagem limiarizada.

Justino [Justino, 2001], acrescenta a virtude dessa característica incorporar um descritor estatístico, o que propicia uma insensibilidade às variações intrapessoais. A Figura 4.9 demonstra tal característica.

4.3.4 Inclinação Axial

Baseado nos estudos de Qi e Hunt [Qi and Hunt, 1995], Justino [Justino, 2001] e Santos [Santos, 2004], essa característica descreve aspectos dinâmicos dos traçados, Figura 4.10.

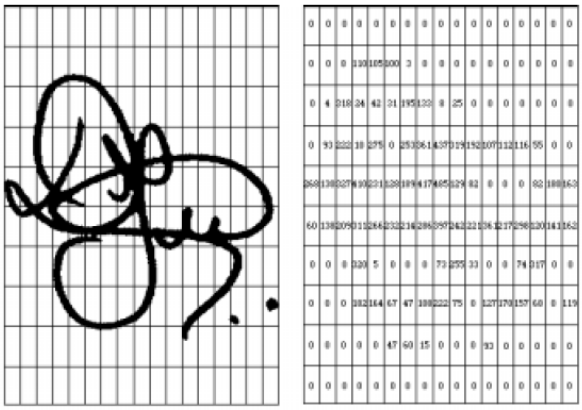


Figura 4.9: Primitiva densidade de *pixels*. Adaptado de [Justino, 2001]

Figura 4.10: Primitiva inclinação axial. Adaptado de [Justino, 2001]

Qi e Hunt [Qi and Hunt, 1995] *appud* Justino [Justino, 2001], dividem em duas etapas o processo para determinar o ângulo de inclinação axial, no qual os resultados das duas etapas são combinados a fim de ter como saída o ângulo global de inclinação axial de uma assinatura. As etapas são:

1. Através das células dos *grids*, é realizada uma análise local sobre os segmentos mais significativos, extraindo o ângulo através do segmento que produziu a máxima projeção.

2. Utiliza-se uma análise global sobre a direção dos segmentos da assinatura, em que o elemento estruturante com o maior número de ocorrências dentro de cada célula dos *grids* representará a inclinação global.

Justino [Justino, 2001], acrescenta que fazer uso de somente um valor angular pode levar à generalização. Isso principalmente em casos de falsificações simuladas. A Figura 4.11 exemplifica o processo de inclinação axial, utilizado por Qi e Hunt [Qi and Hunt, 1995].

4.4 Classificação

A justificativa do SVM é que o mesmo é apropriado para duas classes (genuína e falsificação). Além disso, Justino *et al.* [Justino et al., 2005] demonstra que o classificador SVM produz bons resultados para problemas com verificação de assinaturas *off-line*. Utilizou-se o LIBSVM com a opção de gerar estimação de probabilidades *a posteriori* como saídas.

Sendo um dos objetivos avaliar o impacto que o número de assinaturas de referências influi nas taxas, foram utilizados nos experimentos $Sk = \{3, 5, 7, 9, 11, 13 \text{ e } 15\}$. Através do uso de diferentes números de assinaturas como referência, fica possível estipular até que

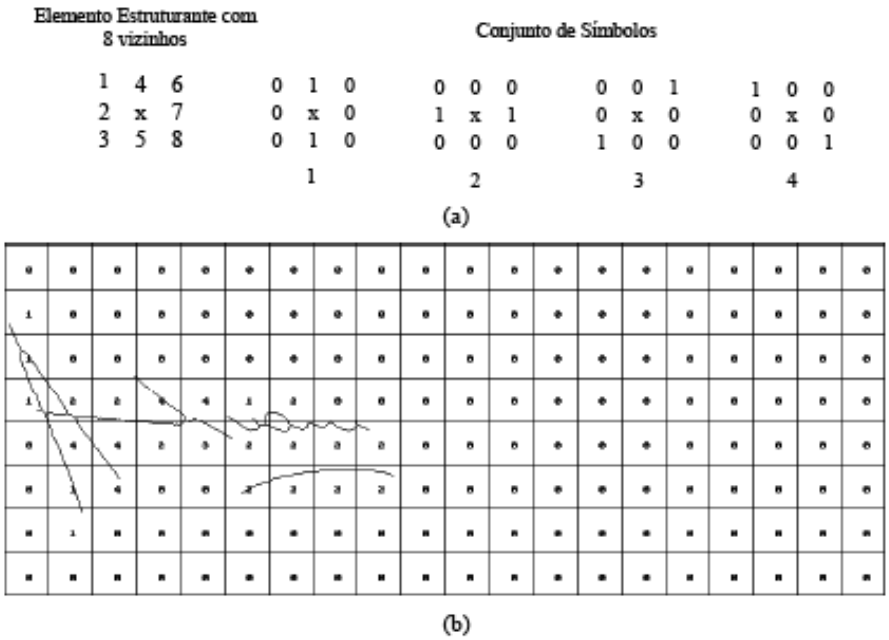


Figura 4.11: Ilustração do processo de extração da primitiva inclinação axial. Adaptado de [Santos, 2004]

ponto a quantidade de assinaturas influencia na taxa de acertos, já que na prática o número de assinaturas é um tanto quanto limitado.

Temos então quatro características extraídas (Distribuição de *Pixels*, Curvatura, Densidade de *Pixels* e Inclinação) e 16 diferentes configurações de *grids*. Para cada combinação (característica $\times$ *grid*) que possuímos como entrada no treinamento, temos um classificador C_i com decisões parciais, isto para cada Sk_x. Dessa forma teremos $Sk_x = \{C_1, C_2, ..., C_p\}$, no qual para *x* teremos (3, 5, 7, 9, 11, 13, 15) e *p* = 64. Em suma para cada número de *Sk* teremos 64 diferentes classificadores.

A taxa de erro global dos classificadores descreve erros entre 9% e 25% em relação ao conjunto de testes, considerando cinco referências. É importante salientar que o SVM foi treinado com amostras provenientes dos 40 autores do conjunto de treinamento. Tais taxas de erros são computadas com base nos 40 autores que não contribuem para o treinamento do classificador de escritor independente. A Figura 4.12 apresenta uma taxa de erro global para a base de classificadores. Observa-se que o conjunto de classificadores formado pela característica distribuição de *pixels* possui os melhores classificadores, enquando o conjunto de característica de curvatura contém classificadores que apresentam taxas de erros bem acima da média.

A Tabela 4.2 apresenta as taxas de erros global e individual, referente ao melhor classificador de cada conjunto de característica, considerando cinco referências. Observa-se as taxas de erro tipo I e erro tipo II sendo apresentadas separadamente.

4.5 Combinando Saídas dos Classificadores

Combinar a saída de classificadores através de algum esquema de fusão é uma técnica bastante usada. Entretanto, geralmente utiliza-se um esquema de votos, no qual através dos rótulos de saída do classificador, é feita uma votação de modo que a classe que tenha mais

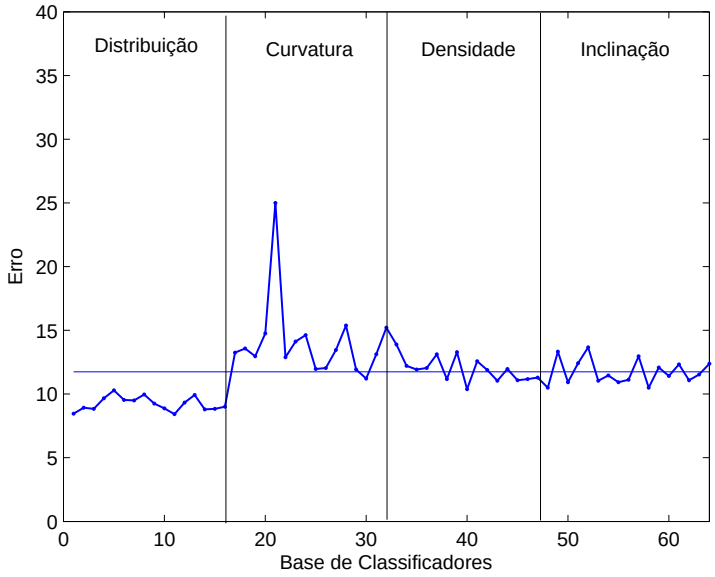


Figura 4.12: Desempenho da base de classificadores.

Tabela 4.2: Melhor classificador de cada conjunto de características referente ao conjunto de testes.

Características	Erro	Erro	Erro Tipo II		
	Global	Tipo I	Simulada	Aleatória	Simples
Distribuição	8,42	18,83	7,50	3,66	3,66
Curvatura	11,19	**27,16**	9,48	**3,32**	4,80
Densidade	10,35	25,32	7,80	3,80	4,48
Inclinação	10,48	**17,48**	10,00	**6,80**	7,64

votos é eleita (genuína ou falsificação). A fim de verificar o impacto que esquemas de fusão proporcionam nas taxas de acertos, propusemos avaliar seis diferentes regras de combinação, sendo elas: Regra do Produto, Soma, Máximo, Mínimo, Média e Mediana.

Nessa etapa, devido ao uso de diferentes números de Sk e este parâmetro tendo impacto no classificador de saída, aplica-se cada esquema de fusão para cada número de Sk, sendo: 3, 5, 7, 9, 11, 13 e 15. Sendo assim, quando utilizado um Sk = 5, obtem-se cinco saídas para cada assinatura questionada (Sq), gerando uma saída para cada esquema de combinação, a qual é um consenso do esquema utilizado. Esse procedimento é aplicado aos 64 classificadores de saída referente ao que cada valor de Sk proporciona.

A Figura 4.13 apresenta um esquema de combinação das saídas de classificadores através de um Sk = 5. Seis diferentes regras de combinação são utilizadas.

A decisão neste caso é definida em função de um limiar. Assim, se o nível de confiança (*score*) dado pelo classificador for maior que 0,5, significa que ele classificou um exemplo como sendo genuíno. Caso essa taxa seja menor que 0,5, a amostra é tida como uma falsificação.

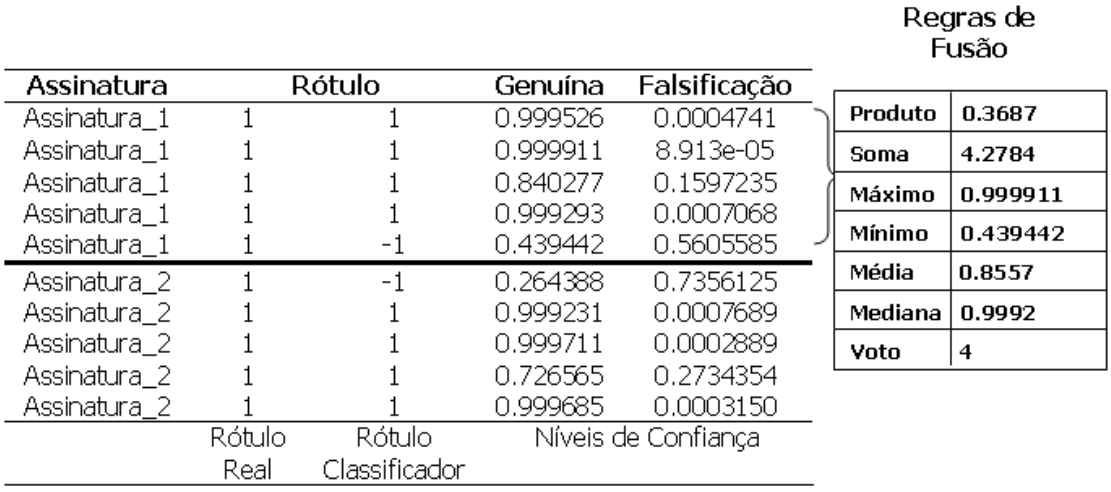


Figura 4.13: Esquema de combinação das saídas dos classificadores utilizando um Sk = 5.

4.6 Agrupamento de Classificadores (*Ensemble de Classificadores*)

Nessa seção apresenta-se a idéia central do trabalho que é agrupar classificadores dinamicamente a fim de que suas decisões combinadas proporcionem melhores taxas. Para isso, utilizaremos esquemas de fusões e Algoritmos Genéticos.

Para realização desses experimentos utilizou-se o software MATLAB em conjunto com a *toolbox* GATool. O MATLAB é um software muito empregado na computação técnica, apresentando alto desempenho e facilidade para trabalhar com diversos tipos de problemas específicos, como reconhecimento de padrões, redes neurais, algoritmos genéticos entre muitos outros.

Diversos experimentos foram realizados com AGs no intuito de identificar configurações que apresentassem melhores resultados. Assim, testes quanto ao tamanho da população, métodos de seleção, taxa de cruzamento, taxa de mutação e critérios de parada foram realizados. Contudo, através destes experimentos conseguimos fixar parâmetros os quais apresentaram o melhor desempenho.

Nos experimentos, um conjunto de 64 classificadores formados a partir das combinações das saídas dos classificadores é tido como entrada (nossa população inicial). Buscou-se então, avaliar o impacto que o número de assinaturas usadas como referência exerce sobre a taxa final. Dessa maneira, para cada número de Sk (3, 5, 7, 9, 11, 13 e 15), foram obtidos seis conjuntos de classificadores (cada um formado por uma regra de combinação das saídas dos classificadores).

Em conjunto com a análise do impacto da formação de conjunto de classificadores agrupados dinamicamente, outro objetivo principal em nosso trabalho é avaliar o desempenho de diferentes funções de aptidão utilizadas no algoritmo genético.

Na busca de construir agrupamentos de classificadores através de AGs, utiliza-se geralmente funções de aptidão que busquem minimizar a taxa de erro global ou também busquem maximizar alguma medida de diversidade. No contexto de verificação de assinaturas, minimizar a taxa de erro global pode não ser ideal, pois a hipótese da distribuição das classes entre as amostras ser constante e relativamente balanceada pode não se manter. Assim, curvas ROC são interessantes devido à propriedade de ser insensível a mudanças de distribuição das classes.

Para isto, três funções de aptidão serão avaliadas, duas delas sendo derivadas de curvas ROC e a terceira da minimização da taxa de erro global.

A primeira é a área abaixo da curva ROC (AUC), sendo o AUC um único valor escalar, que representa o desempenho esperado. Segundo Fawcett [Fawcett, 2006], o AUC possui uma importante propriedade estatística, sendo que o AUC representa a probabilidade de um classificador colocar um exemplo positivo, escolhido aleatoriamente, mais alto na ordenação do que um exemplo negativo. A segunda função objetivo é a maximização do TPR (Taxa de Verdadeiros Positivos) para um determinado FPR (Taxa de Falsos Positivos). A FPR geralmente é uma taxa imposta pela aplicação. Para esses experimentos uma taxa de 10% foi fixada para FPR. Quanto ao cálculo para minimização da taxa de erro global, é apresentado na Equação 4.3. Nesse caso P_I é a probabilidade *a priori* da ocorrência do erro tipo I e P_{II} é a probabilidade *a priori* de ocorrer um erro tipo II. Para o conjunto de testes, consideramos, $P_I = 0,25$ e $P_{II} = 0,75$.

$$\text{Erro Global} = \text{Erro Tipo I } * P_I + \frac{\sum \text{Erro Tipo II}}{3} * P_{II} \tag{4.3}$$

4.7 Cenários Utilizados

Nos experimentos realizados, a base utilizada no processo de formação de agrupamentos com AGs refere-se à base de validação. Dois cenários serão considerados em nosso trabalho, classificados como Cenário I e Cenário II.

No Cenário I, assume-se possuir falsificações. Nesse caso, trabalhamos com assinaturas genuínas e falsificações aleatórias, simples e simulada. Dessa forma, para a formação de agrupamentos de classificadores, três tipos de falsificações são considerados.

Para o Cenário II, procura-se reproduzir experimentos de acordo com casos reais. Casos em que não possui falsificações simples nem simuladas. O que utiliza-se aqui, são falsificações aleatórias, as quais são assinaturas genuínas de outros autores. Assim, para o processo de formação dos agrupamentos de classificadores, somente assinaturas genuínas e falsificações aleatórias são utilizadas.

Através dos dois cenários, pode-se avaliar o impacto na taxa de erros ao trabalharmos com falsificações simples e simuladas e ao não possuí-las. O interessante nesse caso, é que, se as falsificações se tornarem disponíveis num segundo momento, pode-se refazer o processo de agrupamento sem a necessidade de retreinar os classificadores.

4.8 Interpretação dos Resultados

A idéia aqui é a compreensão dos resultados a fim de verificar a relação entre os objetivos iniciais e os resultados obtidos. Nessa etapa avalia-se diversos parâmetros como: aptidão, número de referência (Sk) e uso de diferentes cenários que podem influenciar no problema. O Capítulo 6 descreve detalhadamente essa fase.

Capítulo 5

Experimentos e Resultados

Nesse capítulo será demonstrado os experimentos realizados e os resultados obtidos a partir da metodologia proposta. Um minucioso trabalho de análise dos resultados é realizado nessa etapa a fim de investigar a eficiência em relação aos métodos aqui usados. Dois cenários são utilizados nos experimentos com Algoritmos Genéticos, dentre os quais o primeiro possui respectivamente 2400 assinaturas, sendo 600 genuínas, 600 falsificações aleatórias, 600 falsificações simples e 600 falsificações simuladas, e o segundo cenário com 600 genuínas e 600 falsificações aleatórias.

5.1 Experimentos e Análise em Relação à Combinação da Saída dos Classificadores

Diferentes esquemas de fusão foram utilizados para combinar as saídas dos classificadores. Regras como produto, soma, máximo, mínimo, média e mediana foram avaliadas em nossos experimentos. A idéia é avaliar qual regra pode contribuir para o melhor desempenho, quando utilizadas com intuito de combinar as saídas dos classificadores.

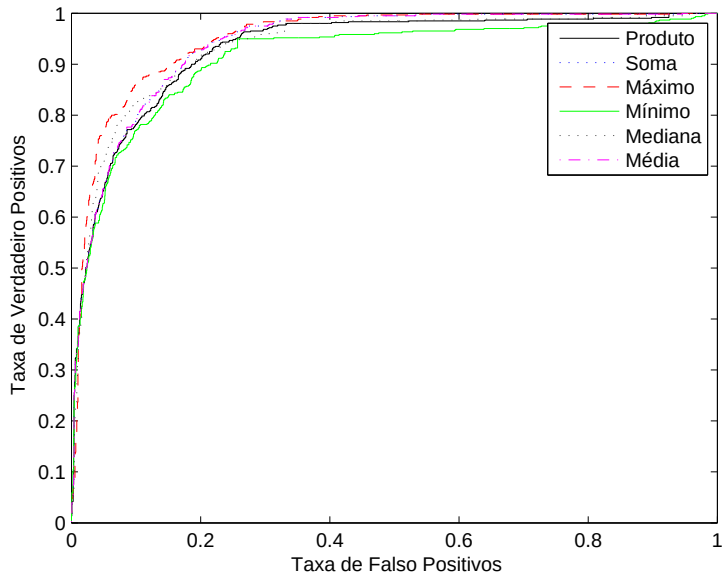
Avaliando os resultados finais proporcionados através dos diferentes esquemas de fusão, observa-se que a regra do máximo apresenta desempenho superior as demais. Verifica-se também que a regra do mínimo foi a que apresentou os piores resultados. Esquemas como da soma, mediana e produto apresentaram bons resultados em diversos casos.

Demonstra-se a seguir alguns dos experimentos realizados com a combinação das saídas dos classificadores. Para isso dois casos foram avaliados isoladamente. No primeiro, utiliza-se 3 assinaturas como referência ($Sk = 3$) e no segundo, 15 assinaturas ($Sk = 15$). O classificador refere-se à característica de Distribuição de *pixels* com uma divisão de *grids* de 4×5, (primeiro classificador da característica Distribuição).

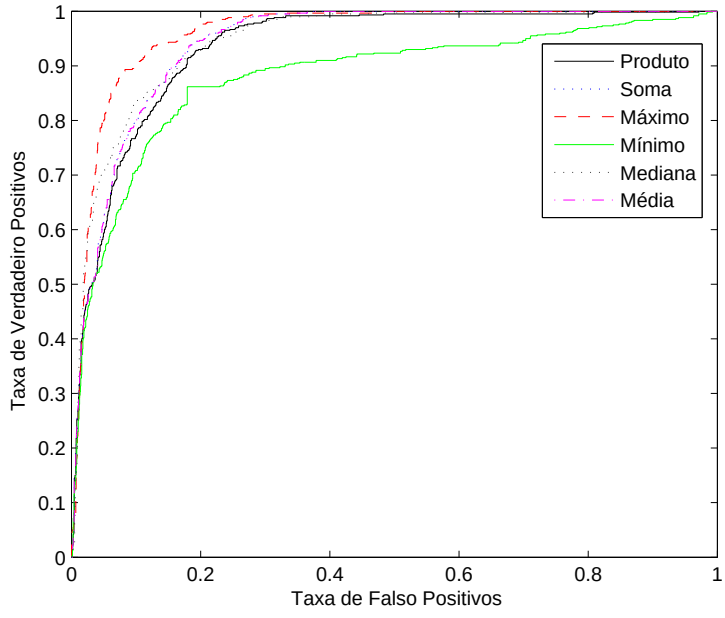
A Tabela 5.1 demonstra as taxas de erro global e a AUC dos dois casos. Observa-se nesses dois casos que a regra do máximo apresentou resultados superiores aos demais. O mesmo foi obervado em outros números de referência e demais características.

A Figura 5.1 demonstra as curvas ROC referentes a utilização dos diversos esquemas de combinação. Para essa base de dados, fica claro a superioridade do esquema do máximo para se combinar saídas de classificadores.

Nos experimentos para construir agrupamentos de classificadores através de AGs, foram testados conjuntos de classificadores formados por todas as regras de fusão. Contudo como o



(a)



(b)

Figura 5.1: Avaliação de desempenho quanto aos esquemas de combinação de classificadores. (a) $Sk = 3$ e (b) $Sk = 15$.

Tabela 5.1: Avaliação do uso de diferentes esquemas de fusão para combinação das saídas de classificadores.

Esquema de	Erro Global		AUC	
Fusão	*Sk*=3	*Sk*=15	*Sk*=3	*Sk*=15
Produto	11,84	12,46	0,9308	0,9342
Soma	11,75	12,00	0,9405	0,9423
Máximo	**9,30**	**8,30**	**0,9501**	**0,9606**
Mínimo	12,17	14,50	0,9133	0,8800
Mediana	10,80	10,92	0,9348	0,9443
Média	15,75	12,00	0,9405	0,9423

resultado final, após o agrupamento formado, em todos os casos apresentou as melhores taxas quando utilizamos o conjunto formadoo pela regra do máximo, todos os testes e resultados a serem apresentados a seguir são realizados com base nesse conjunto.

5.2 Experimentos e Análise em Relação às Funções Objetivo

Um importante estudo a ser realizado refere-se ao impacto proporcionado pela aptidão utilizada no AG. Sabe-se que a função objetivo é inerente ao problema em questão. Neste trabalho avaliou-se três diferentes funções de aptidão, taxa de erro global, AUC e a TPR para FPR fixada em 10%.

Os resultados apresentados na Tabela 5.2, referem-se às melhores taxas proporcionadas pelos agrupamentos de classificadores. Várias regras de fusão foram testadas. As taxas apresentadas a seguir referem-se à validação do Cenário I, no qual consideramos as assinaturas genuínas e três tipos de falsificações, lembrando que tais classificadores são resultados do esquema de combinação do máximo.

Tabela 5.2: Taxa de erro global e AUC das diferentes funções de aptidão utilizadas, Cenário I.

$Sk \rightarrow$	$Sk = 3$		$Sk = 9$		$Sk = 15$	
Aptidão $\downarrow$	Erro Global	AUC	Erro Global	AUC	Erro Global	AUC
Erro Global	7,13	0,9676	6,13	0,9732	5,67	0,9807
AUC	8,09	0,9710	6,59	0,9803	5,92	0,9819
FPR fixada 10%	8,05	0,9688	6,84	0,9788	6,00	0,9804

As Figuras 5.2, 5.3 e 5.4 apresentam as curvas ROC referentes às três funções objetivo utilizadas neste trabalho. Avaliando diferentes conjuntos de referências pode-se observar que quanto menor for o tamanho do conjunto de referencia maior será o impacto da função objetivo durante a otimização.

Na Figura 5.2, em que $Sk = 3$, observa-se que a curva mais homogênea do gráfico refere-se ao agrupamento que maximiza a AUC. Fica claro também que a maximização do ponto fixo da taxa de falso positivos em 10% realmente melhorou o resultado naquele ponto. O pior ocorre quando miminizamos a taxa de erro global. Isto pode ser entendido pelo fato da taxa de erro global ser mais sensível à distribuição das classes.

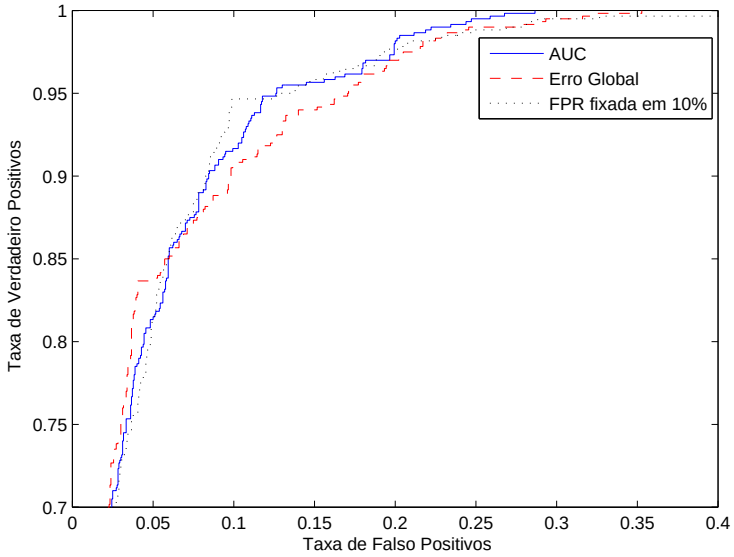


Figura 5.2: Comparação entre as três funções objetivos consideradas neste trabalho. *Sk* = 3. Cenário I.

Observando as Figuras 5.3 e 5.4 consegue-se concluir que conforme aumentamos o tamanho do conjunto de referencia (*Sk*) há uma minimização do impacto das funções objetivos, ou seja, indiferente da aptidão utilizada consegue-se resultados muito próximos, isto quando se tem um alto número de referências, o que sabemos ser raramente disponível. Com isto, acreditamos que a função objetivo derivada da curva ROC é mais adequada no contexto de verificação de assinaturas.

As curvas ROC apresentadas referem-se a testes levando em consideração a validação do Cenário I, no qual assinaturas genuínas, falsificações aleatórias, simples e simuladas são utilizadas. Avaliaremos a seguir o impacto das mesmas funções objetivos com o Cenário II. A Tabela 5.3 apresenta as taxas de erro global e AUC relativas às funções objetivo referente ao Cenário II.

Tabela 5.3: Erro Global e AUC das diferentes funções objetivos utilizadas, Cenário II.

Sk →	*Sk* = 3		*Sk* = 9		*Sk* = 15	
Aptidão ↓	Taxa Rec.	AUC	Taxa Rec.	AUC	Taxa Rec.	AUC
Erro Global	8,88	0,9694	7,17	0,9771	7,50	0,9803
AUC	7,88	0,9604	6,34	0,9789	7,96	0,9799
TPR fixada em 10%	8,00	0,9673	6,80	0,9775	7,80	0,9808

As Figuras 5.5, 5.6 e 5.7 apresentam respectivamente as curvas ROC proporcionadas utilizando *Sk* = 3, 9 e 15, respectivamente. Observando essas curvas, e realizando um estudo comparativo com as curvas apresentadas quanto ao Cenário I, percebe-se que do mesmo modo como no Cenário I, com o aumento do número de referências, minimiza-se o impacto da função objetivo.

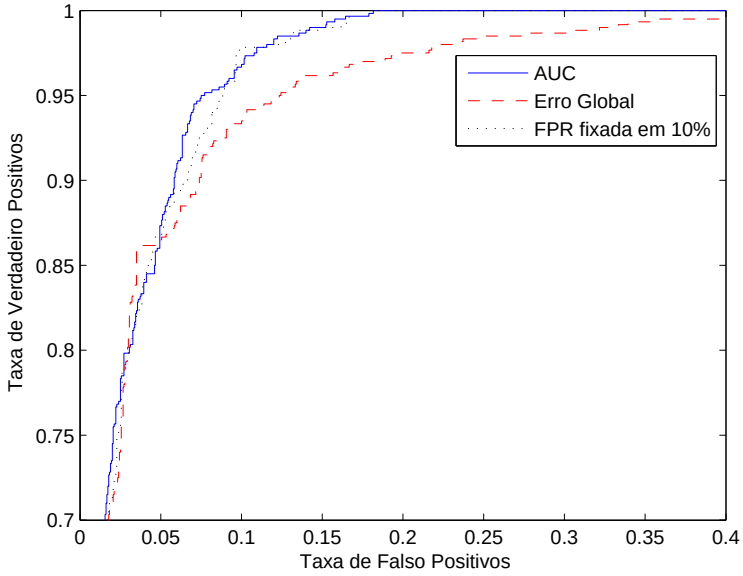


Figura 5.3: Comparação entre as três funções objetivos consideradas neste trabalho. *Sk* = 9, Cenário I.

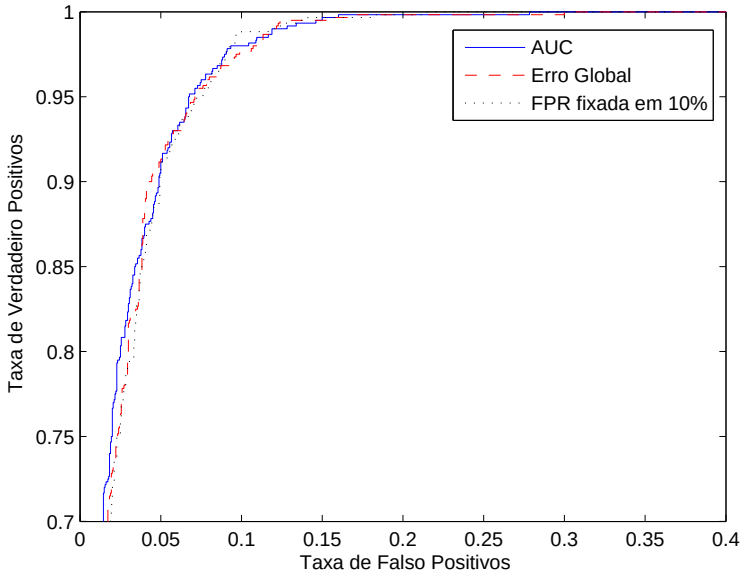


Figura 5.4: Comparação entre as três funções objetivos consideradas neste trabalho. *Sk* = 15, Cenário I.

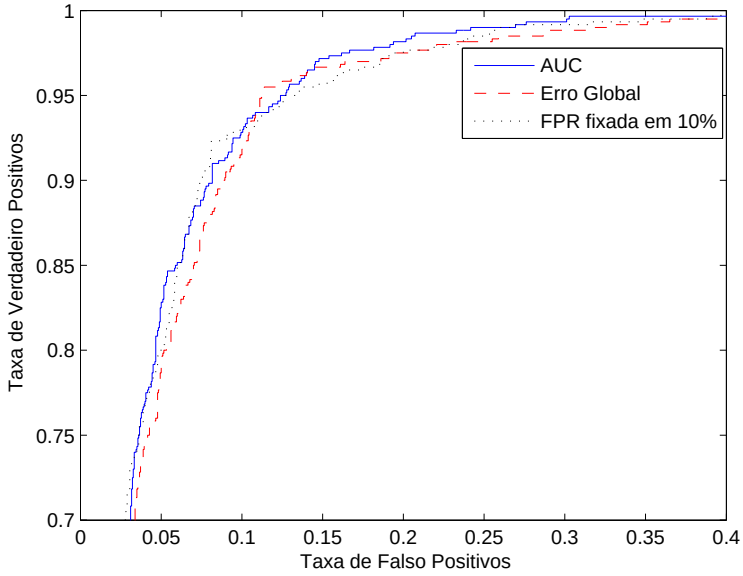


Figura 5.5: Comparação entre as três funções objetivos consideradas nesse trabalho. *Sk* = 3, Cenário II.

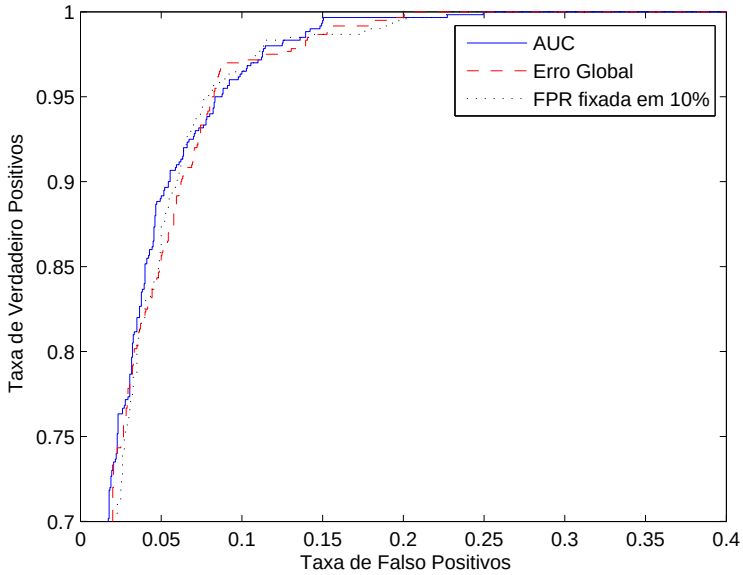


Figura 5.6: Comparação entre as três funções objetivos consideradas nesse trabalho. *Sk* = 9, Cenário II.

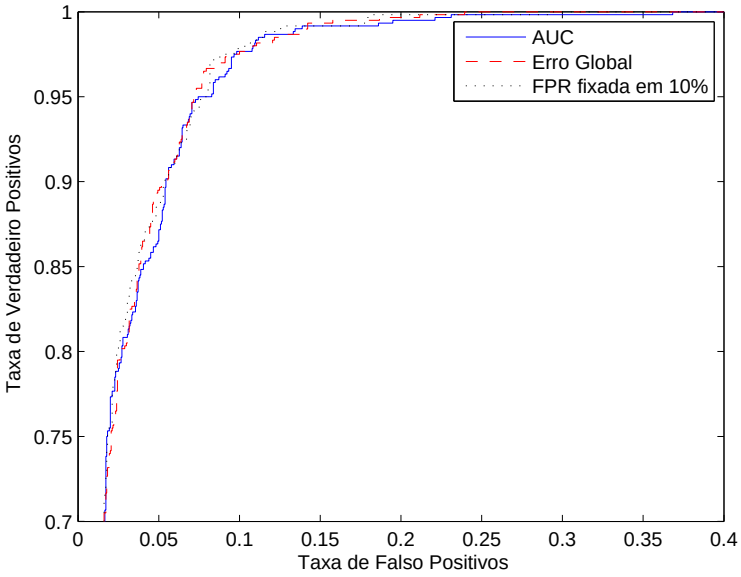


Figura 5.7: Comparação entre as três funções objetivos consideradas nesse trabalho. *Sk* = 15, Cenário II.

5.3 Experimentos e Análise em Relação ao Tamanho do Conjunto de Referências

Uma questão importante da abordagem proposta nesse trabalho é o tamanho do conjunto de referências (*Sk*) e o impacto quanto a confiabilidade do sistema de verificação de assinaturas. Entretanto, aplicações diferentes podem ter prioridades diferentes. É evidente que um sistema confiável é aquele que reduz simultaneamente os erros do tipo I e tipo II. Aplicações reais podem otimizar um ou outro tipo de erro, isso depende muito do nicho de aplicação do sistema. Para determinadas situações, reduzir erros do tipo II é muito mais importante que reduzir erros do tipo I, ou vice-versa.

Nesse trabalho, assume-se que a confiabilidade do sistema está em ser resistente a falsificações, sendo necessário, pois, reduzir o máximo possível do erro tipo II.

As Tabelas 5.4, 5.5 e 5.6 apresentam as melhores taxas proporcionadas pelos agrupamentos de classificadores. Diferentes funções objetivos com diferentes tamanhos do conjunto de referência (*Sk*) são avaliados. Os dados dessas tabelas refem-se a experimentos realizados com o Cenário I. Observa-se nas Tabelas 5.4, 5.5 e 5.6 verifica-se em que condições o sistema é mais resistente a falsificações.

Em princípio, com os experimentos realizados, observa-se a importância da utilização de falsificações simples e simuladas no processo de otimização do agrupamento. O aspecto interessante de possuir classificadores universais baseados na dissimilaridade é que se falsificações simples e simuladas estiverem disponíveis, elas podem encontrar conjuntos mais confiáveis sem o retreinamento dos classificadores.

Avaliando os resultados das Tabelas 5.4, 5.5 e 5.6, verifica-se que ao utilizar o Cenário I e indiferente da função objetivo utilizada, obtém-se melhores resultados. As melhores taxas alcançadas em relação à minimização do erro tipo II, foram conseguidas quando utilizamos

Sk	Erro Global	Erro	Erro Tipo II		
		Tipo I	Simulada	Aleatória	Simples
3	8,06	14,32	8,64	4,48	4,80
5	7,27	21,08	**3,80**	**2,00**	**1,48**
7	6,65	12,48	6,48	3,64	4,00
9	6,46	**7,32**	8,32	5,32	4,88
11	6,74	9,16	8,32	5,00	4,48
13	6,36	11,00	7,00	3,64	3,80
15	5,09	8,32	7,16	3,80	4,32

Tabela 5.4: Resultados dos experimentos utilizando a taxa de erro global como função objetivo, Cenário I.

Sk	Erro Global	Erro	Erro Tipo II		
		Tipo I	Simulada	Aleatória	Simples
3	7,12	16,32	6,16	3,00	3,00
5	6,61	17,64	**4,64**	**2,00**	**2,16**
7	6,03	11,16	6,48	3,16	3,32
9	6,11	14,00	5,00	2,80	2,64
11	6,36	14,00	6,00	2,80	2,64
13	6,20	**9,16**	7,48	4,16	4,00
15	5,65	10,16	6,48	3,16	2,80

Tabela 5.5: Resultados dos experimentos utilizando a AUC como função objetivo, Cenário I.

Sk	Erro Global	Erro	Erro Tipo II		
		Tipo I	Simulada	Aleatória	Simples
3	8,02	13,80	8,16	5,32	4,80
5	7,57	18,00	**6,16**	**2,80**	**3,32**
7	7,02	14,32	6,48	3,48	3,80
9	6,81	13,80	6,00	3,64	3,80
11	7,12	10,16	9,16	4,00	5,16
13	7,03	10,16	9,32	4,16	4,48
15	5,99	**9,00**	7,48	3,48	4,00

Tabela 5.6: Resultados dos experimentos utilizando o FPR fixada em 10% como função objetivo, Cenário I.

cinco assinaturas como referência (*Sk* = 5). Entretanto, a otimização da taxa global alcança melhores resultados ao utilizar um número mais alto de referências (15).

Fica claro a partir das experiências realizadas que aumentar o tamanho do conjunto de referência *(Sk)* não necessariamente reduz erro tipo I, mas, geralmente, reduz o erro global.

Uma análise quanto as curvas ROC de agrupamentos pode ser observadas nas Figuras 5.8, 5.9 e 5.10. Indiferente da função objetivo utilizada, nota-se um aumento na área da curva ROC quando aumentado o número de assinaturas de referência.

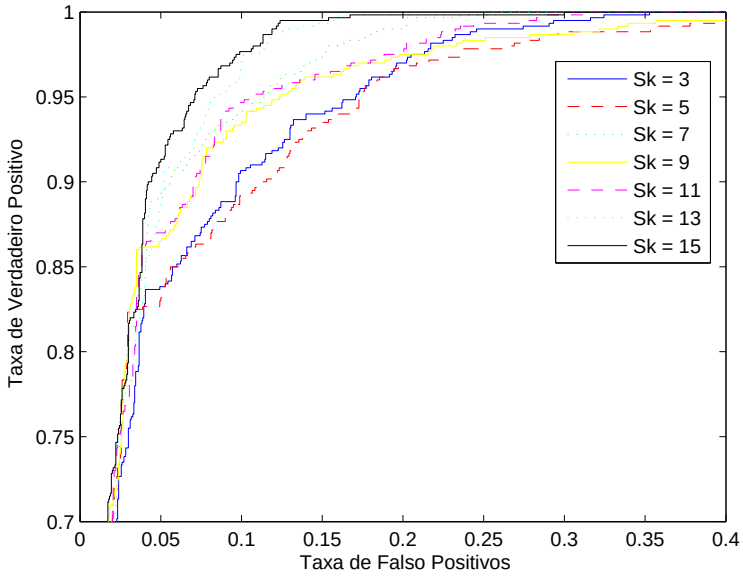


Figura 5.8: Comparação entre diferentes números de (*Sk*) considerados nesse trabalho, usando como função objetivo a taxa de erro global, Cenário I.

Se não existirem falsificações do tipo simples e simulada para a otimização, o número de referências para alcançar um resultado semelhante deve ser aumentado. Veja as Tabelas 5.7, 5.8 e 5.9.

Sk	Erro Global	Erro	Erro Tipo II		
		Tipo I	Simulada	Aleatória	Simples
3	8,85	20,32	6,80	3,80	4,48
5	8,81	16,16	7,64	4,80	6,64
7	7,82	12,32	9,00	4,64	5,32
9	7,15	10,80	7,64	5,16	5,00
11	7,19	**9,64**	9,48	4,48	5,16
13	7,54	17,00	7,00	**3,00**	3,16
15	6,28	11,32	**6,48**	4,32	3,00

Tabela 5.7: Resultados dos testes utilizando a taxa de erro global como função objetivo, Cenário II.

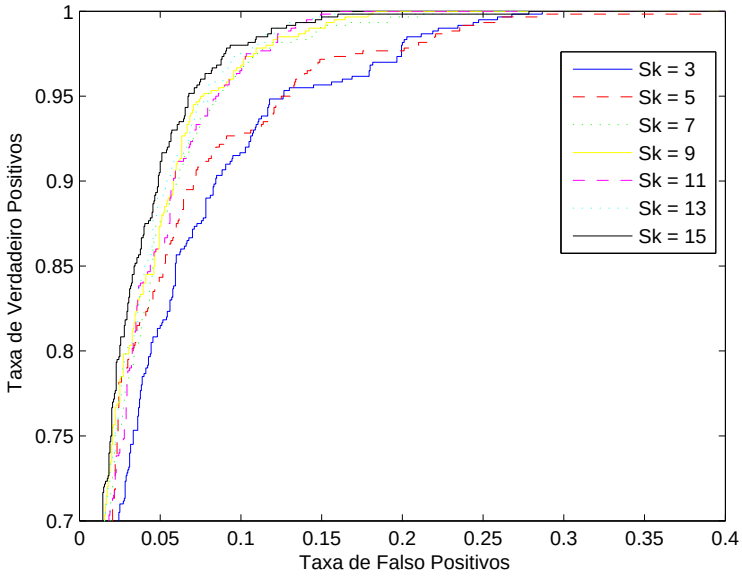


Figura 5.9: Comparação entre diferentes números de (*Sk*) considerados nesse trabalho, usando como função objetivo a AUC, Cenário I.

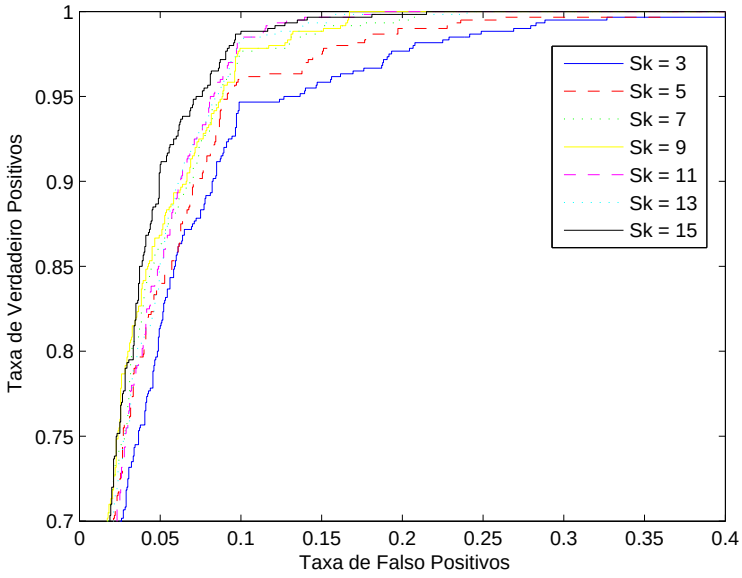


Figura 5.10: Comparação entre diferentes números de (*Sk*) considerados nesse trabalho, usando como função objetivo a TPR fixada em 10%, Cenário I.

Sk	Erro Global	Erro	Erro Tipo II		
		Tipo I	Simulada	Aleatória	Simples
3	7,86	15,32	7,48	4,16	4,48
5	7,32	11,32	8,00	4,48	5,48
7	6,32	11,32	5,00	4,16	4,80
9	7,04	10,00	7,00	5,00	6,16
11	7,19	17,64	**4,32**	**3,32**	**3,48**
13	6,73	**7,32**	7,80	5,32	6,48
15	6,48	9,16	6,64	4,80	5,32

Tabela 5.8: Resultados dos testes utilizando a AUC como função objetivo, Cenário II.

Sk	Taxa Erro	Erro	Erro Tipo II		
		Tipo I	Simulada	Aleatória	Simples
3	7,99	7,64	10,00	7,16	7,16
5	7,78	10,80	9,00	5,32	6,00
7	7,28	13,00	7,16	4,64	4,32
9	6,80	8,00	8,16	5,32	5,64
11	6,88	**5,16**	9,48	6,64	6,16
13	6,98	15,16	5,64	3,32	3,80
15	6,34	13,64	**5,16**	**3,16**	**3,32**

Tabela 5.9: Resultados dos testes utilizando a FPR fixada em 10% como função objetivo, Cenário II.

5.4 Análise quanto aos Classificadores Selecionados

Um ponto importante em relação à formação de agrupamentos de classificadores consiste em analisar os classificadores selecionados. Em teoria, um bom agrupamento é composto por bons classificadores, porém não necessariamente por excelentes classificadores.

Tal comportamento pode ser observado em nossos experimentos, nos quais vários classificadores com bom desempenho não foram selecionados e outros com desempenho abaixo da média foram escolhidos para fazer parte do conjunto.

As Figuras 5.11 e 5.12 referem-se aos classificadores escolhidos para formar o agrupamento utilizando os Cenários I e II, respectivamente. Nos dois casos as três funções objetivas foram avaliadas.

Conforme apresentam as Figuras 5.11 e 5.12, os classificadores treinados com a característica inclinação são selecionados mais frequentemente. Contudo a Figura 4.12(Capítulo 4, Seção 4.4), mostra que tais classificadores não possuem mais poder discriminativo que características como a Distribuição ou a Densidade. O fato desses serem selecionados com maiores freqüências deve-se ao fornecimento de informações complementares que, em união com outras características proporcionam um melhor resultado.

Na Tabela 4.2 (Capítulo 4, Seção 4.4), verifica-se que o erro global fornecido pelos classificadores baseado na Distribuição dos *pixels* são menores que os proporcionados pela característica Inclinação. Isto nos prova que classificadores não precisam ser excelentes, entretanto eles devem discordar, tanto quanto possível, em casos de dificuldade.

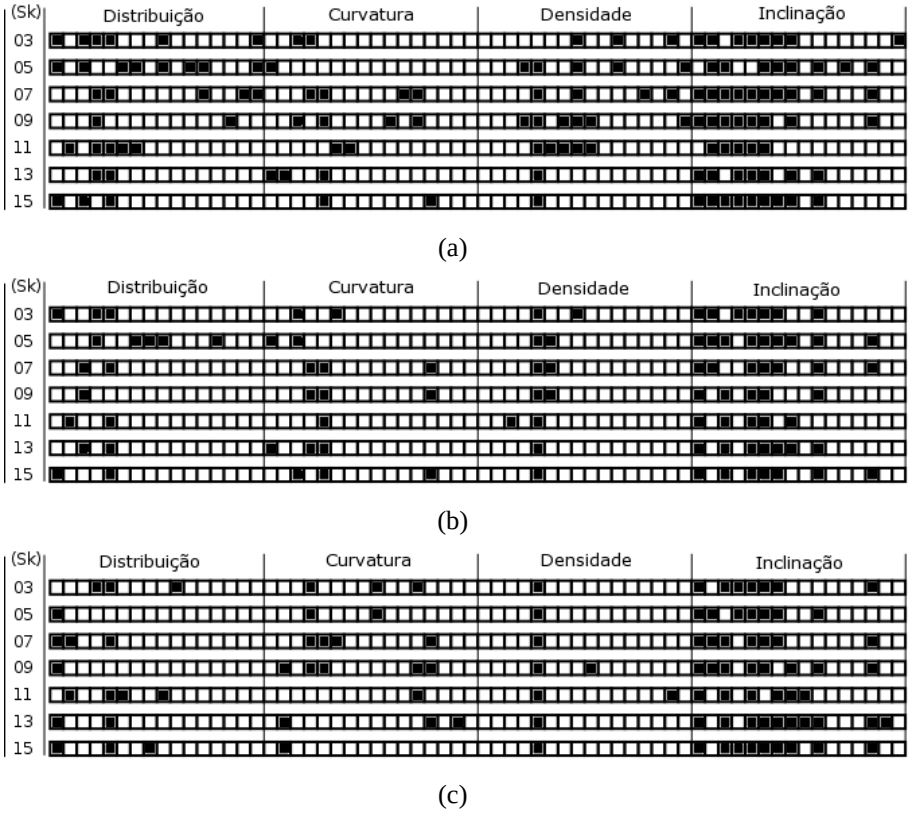


Figura 5.11: Classificadores selecionados para compor o agrupamento utilizando o Cenário I. Aptidão: (a) Erro Global, (b) AUC, (c) FPR fixada em 10%.

As Figuras 5.11 e 5.12, mostram que o conjunto de características da Curvatura desempenha um importante papel no agrupamento de classificadores. Pois até os classificadores mais fracos são frequentemente selecionados. Como exemplo, pode-se citar o quinto classificador referente à característica curvatura que, com uma divisão de *grids* de 5×5 apresentou uma taxa de erro de 25% e foi selecionado diversas vezes para compor o agrupamento. Com isso certifica-se o argumento que a característica de curvatura (curvas Bezier), apresentam informações necessárias complementares para outros conjuntos de características utilizados neste trabalho.

Os classificadores baseados na curvatura são o segundo conjunto de características mais selecionados para formar o agrupamento.

5.5 Avaliação quanto aos Esquemas de Fusão usados com AGs

Avaliou-se o impacto de diferentes regras de fusão na seleção de classificadores para formar o agrupamento dos classificadores. Neste caso, o esquema que apresentou melhor desempenho foi a soma. Os resultados apresentados pela mediana e pelo produto ficaram próximos.

A Figura 5.13 apresenta os resultados utilizando as saídas proporcionadas pela regra do máximo como esquema para se combinar as saídas dos classificadores e cinco assinaturas

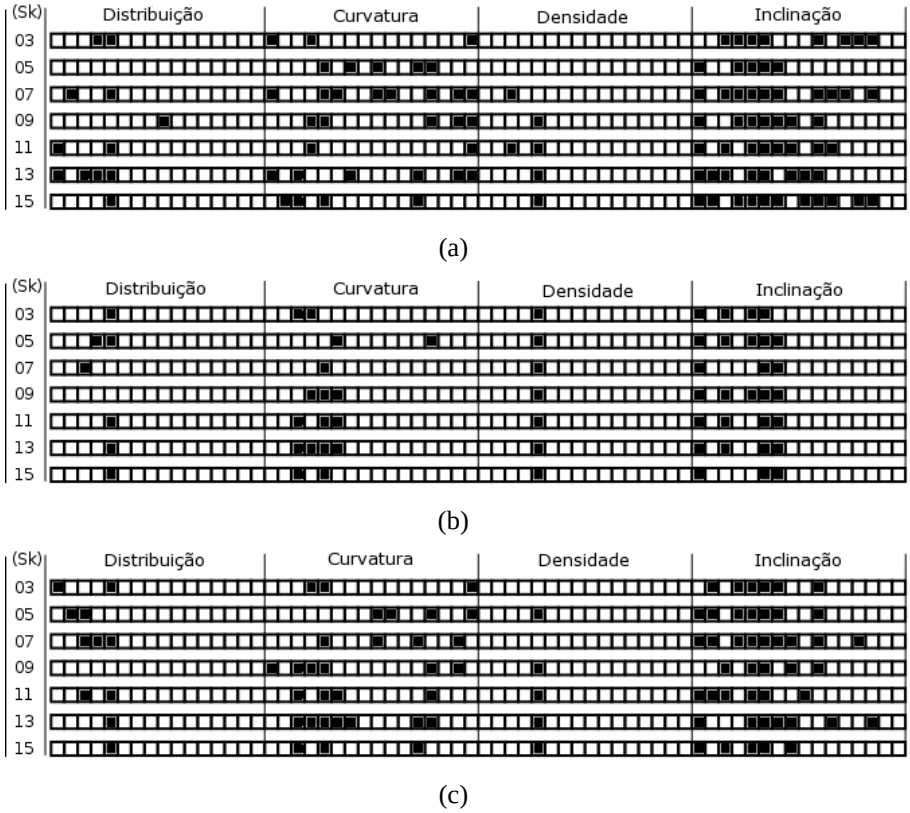


Figura 5.12: Classificadores selecionados para compor o agrupamento utilizando o Cenário II. Aptidão: (a) Erro Global, (b) AUC, (c) TPF fixada em 10%.

como referência (*Sk =5*). Avaliou-se os dois cenários em estudo (Cenário I e Cenário II). Cada experimento foi replicado 10 vezes para avaliar a capacidade de reprodução, sendo os resultados uma média das 10 replicações.

A partir da Figura 5.13 verifica-se o impacto que as regras de fusão podem ter na formação de agrupamento de classificadores. Em geral, percebe-se que a regra do máximo não obteve um desempenho considerável para esses casos. Contudo, é a regra que seleciona menos classificadores para formar o agrupamento. Ao contrário temos a regra do mínimo que seleciona um grande número de classificadores para compor o agrupamento.

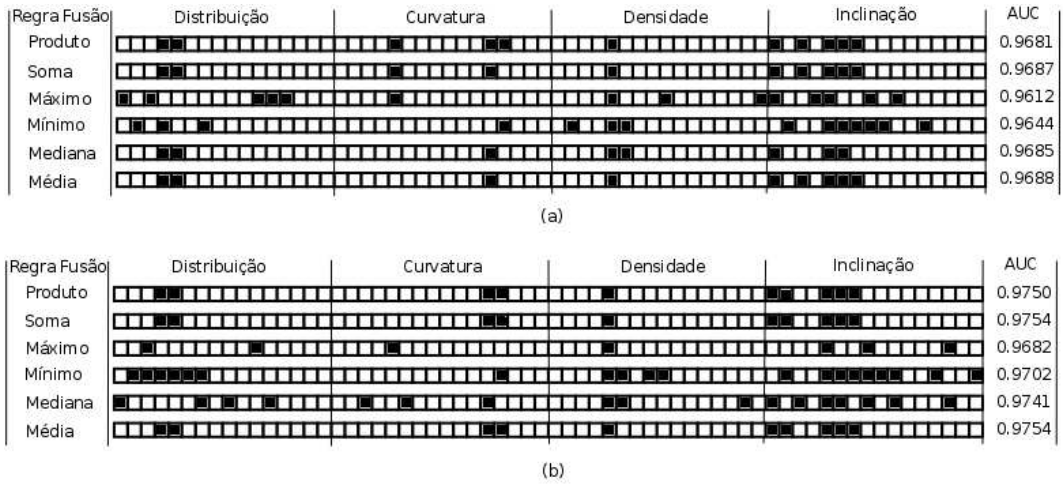


Figura 5.13: Classificadores selecionados para compor o agrupamento utilizando a AUC como aptidão e conjunto de validação: (a) Cenário II, (b) Cenário I

Capítulo 6

Conclusões

A partir dos resultados obtidos e apresentados nos Capítulos anteriores foi possível fazer uma avaliação de desempenho em relação à estratégia do uso de agrupamento de classificadores na verificação de assinaturas *off-line*. Ao analizar os resultados pode-se concluir que através de agrupamento de classificadores consegue-se melhorar a confiabilidade em relação ao desempenho de sistemas de verificação de assinaturas *off-line*.

Após todos os experimentos realizados, como combinação das saídas dos classificadores, algoritmos genéticos, seleção de classificadores para compor o agrupamento, e a avaliação dos resultados quanto aos dois cenários, identificou-se:

- Que o esquema baseado na dissimilaridade, no qual novos usuários podem ser acrescentados à base sem o retreinamento dos classificadores, permitindo a criação de um classificador global, é de grande utilidade. Isso porque, se as falsificações estão disponíveis para alguns escritores que não participaram do processo de treinamento, tais amostras podem ser usadas no aprimoramento do sistema, formando um excelente agrupamento de classificadores;

- Para essa base de dados, o uso da regra do máximo para combinar as saídas dos classificadores apresentou resultados superiores às demais regras. Entretanto, a mesma não apresentou bons resultados quando utilizada no AG para formação do agrupamento de classificadores;

- Quanto aos testes realizados levando em conta duas situações (Cenário I assinaturas genuínas e falsificações aleatórias, simples e simuladas; Cenário II assinaturas genuínas e falsificações aleatórias), observa-se que os agrupamentos baseados em características grafométricas são muito eficientes e podem reduzir consideravelmente o erro tipo II (falsa aceitação). Verificou-se através dos experimentos que, ao possuir falsificações disponíveis, mesmo em número limitado, consegue-se um sistema mais robusto em relação a detecção de falsificações;

- Através de testes com funções objetivos constatou-se que: 1) Ao utilizar um pequeno número de imagens de assinaturas como referência (*Sk*), a AUC utilizada como aptidão apresentou em geral melhores resultados que as outras. 2) Para um alto número de assinaturas de referência, (Sk = 15), os resultados das funções objetivos ficam muito próximos um dos outros;

- Em relação aos classificadores selecionados para compor o agrupamento de classificadores, percebe-se um maior número de classificadores selecionados referente às características de inclinação e curvatura. Sendo estas características pseudo-dinâmicas, pode-se concluir que para a formação dos agrupamentos, características pseudo-dinâmicas possuem um maior desempenho que características estáticas. Contudo, observando a Tabela 4.2 (características: curvatura e inclinação) observa-se que as taxas para erro tipo I e para falsificação simulada são inversamente proporcionais. Tem-se a curvatura apresentando a maior taxa de erro tipo I e a característica inclinação apresentando a menor. Para as taxas de erro referente a falsificação simulada, tem-se a inclinação apresentando a mais alta, e a curvatura a mais baixa taxa. Verifica-se então que tais características formando agrupamentos apresentam um equilíbrio nas taxas de erros.

Em suma, ao avaliar os objetivos propostos e os resultados alcançados, pode-se concluir que grande parte deles foram satisfatórios, pois:

- Como apresentado, alcançamos interessantes resultados quanto à redução dos erros tipo I e II, ver Capítulo , Tabelas 5.4, 5.5 e 5.6;
- Com relação à formação de agrupamento através de classificadores, percebemos que estes foram de excelente valia, pois o taxa de erro foi reduzida e para tal menos utilizou-se menos da metade dos classificadores para compor o agrupamento.
- Quanto aos cenários e número de referências avaliados, observou-se a relação existente entre número de assinaturas usadas como referência e ao possuirmos ou não falsificações do tipo simples, simulada;
- Estudou-se também o impacto das funções de aptidão, nas quais curvas ROC apresentaram melhor desempenho em relação as outras;
- Por fim, foi possível observar que classificadores derivados de características pseudo-dinâmicas (curvatura e inclinação) foram selecionados mais freqüentemente para compor o agrupamento.

6.1 Trabalhos Futuros

Como trabalhos futuros, pretende-se:

- Avaliar o impacto ao utilizar assinaturas genuínas e falsificações simuladas para formar agrupamentos através de algoritmos genéticos;
- Utilizar novas características, conseguindo assim um maior número de classificadores para experimentos com agrupamentos de classificadores.

Referências Bibliográficas

[Ammar, 1991] Ammar, M. (1991). Progress in verification of skillfully simulated handwritten signatures. *International Journal of Pattern Recognition and Artificial Intelligence*, 1(2):337–351.

[Armand et al., 2006] Armand, S., Blumenstein, M., and Muthukkumarasamy, V. (2006). Off-line signature verification using the enhanced modified direction feature and neural-based classification. In *International Joint Conference on Neural Networks*, pages 684–691.

[Bäck, 1996] Bäck, T. (1996). *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press, Oxford, UK.

[Bajaj and Chaudhury, 1997] Bajaj, R. and Chaudhury, S. (1997). Signature verification using multiple neural classifiers. *Pattern Recognition*, 30(1):1–7.

[Baltzakis and Papamarkos, 2001] Baltzakis, H. and Papamarkos, N. (2001). A new signature verification technique based on a two-stage neural network classifier. *Engineering Applications of Artificial Inteligence*, 14:95–103.

[Batista et al., 2007] Batista, L., Rivard, D., Sabourin, R., Granger, E., and Maupin, P. (2007). *State Of The Art In Off-Line Signature Verification*. Pattern Recognition Technologies and Applications: Recent Advances.

[Bernardini, 2006] Bernardini, F. C. (2006). *Combinação de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéticos*. PhD thesis, Instituto de Ciências Matemáticas e de Computação (ICMC).

[Bertolini et al., 2008] Bertolini, D., Oliveira, L. S., Justino, E., and Sabourin, R. (2008). Ensemble of classifiers for off-line signature verification. In *IEEE International Conference on Systems, Man and Cybernetics (SMC2008)*. Artigo Aceito.

[Britto et al., 2001] Britto, A. S., de Almendra Freitas, C. O., Justino, E. J. R., Borges, D. L., Facon, J., Bortolozzi, F., and Sabourin, R. (2001). Técnicas em processamento e análise de documentos manuscritos. *RITA*, 8(2):47–68.

[Burges, 1998] Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

[Cardot et al., 1994] Cardot, H., Revenu, M., Victorri, B., and Revillet, M.-J. (1994). A static signature verification system based on a cooperating neural networks architecture. *IJPRAI*, 8(3):679–692.

[Coetzer et al., 2006] Coetzer, H., Herbst, B., and Du Preez, J. (2006). Off-line signature verification: A comparison between human and machine performance. pages 481–485. IAPR Publishers.

[Coetzer, 2005] Coetzer, J. (2005). *Off-line Signature Verification*. PhD thesis, University of Stellenbosh.

[Deng et al., 1999] Deng, P. S., Liao, H.-Y. M., Ho, C.-W., and Tyan, H.-R. (1999). Wavelet-based off-line handwritten signature verification. *Comput. Vis. Image Underst.*, 76(3):173–190.

[Dietterich, 2000] Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15.

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley Interscience, 2 edition.

[El-Yacoubi et al., 2000] El-Yacoubi, A., Justino, E., Sabourin, R., and Bortolozzi, F. (2000). Off-line signature verification using hmms and cross-validation. *IEEE Workshop on Neural Networks for Signal Processing*, pages 859–868.

[Fang et al., 2003] Fang, B., Leung, C. H., Tang, Y. Y., Tse, K. W., Kwok, P. C. K., and Wong, Y. K. (2003). Off-line signature verification by the tracking of feature and stroke positions. *Pattern Recognition*, 36(1):91–101.

[Fang et al., 2001] Fang, B., Wang, Y. Y., Leung, C. H., Tse, K. W., Tang, Y. Y., Kwok, P. C. K., and Wong, Y. K. (2001). Offline signature verification by the analysis of cursive strokes. *IJPRAI*, 15(4):659–673.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874.

[Guo et al., 1997] Guo, J. K., Doermann, D. S., and Rosenfeld, A. (1997). Local correspondence for detecting random forgeries. In *ICDAR 97, International Conference on Document Analysis and Recognition*, pages 319–323, Washington, DC, USA. IEEE Computer Society.

[Hansen and Salamon, 1990] Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001.

[Holland, 1992] Holland, J. H. (1992). *Adaptation in natural and artificial systems*. MIT Press, Cambridge, MA, USA.

[Huang and Yan, 2002] Huang, K. and Yan, H. (2002). Off-line signature verification using structural feature correspondence. *Pattern Recognition*, 35(11):2467–2477.

[Impedovo and Pirlo, 2008] Impedovo, D. and Pirlo, G. (2008). Automatic signature verification: The state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 38:609–635.

[Ismail and Gad, 2000] Ismail, M. A. and Gad, S. (2000). Off-line arabic signature recognition and verification. *Pattern Recognition*, 33(10):1727–1740.

[Justino, 2001] Justino, E. (2001). *O Grafismo e os Modelos Escondidos de Markov na Verificação Automática de Assinaturas*. PhD thesis, Pontíficia Universidade Católica do Paraná.

[Justino et al., 2001] Justino, E. J. R., Bortolozzi, F., and Sabourin, R. (2001). Offline signature verification using hmm for random. In *ICDAR 2001, International Conference on Document Analysis and Recognition*, pages 1031–1034.

[Justino et al., 2005] Justino, E. J. R., Bortolozzi, F., and Sabourin, R. (2005). A comparison of svm and hmm classifiers in the off-line signature verification. *Pattern Recogn. Lett.*, 26(9):1377–1385.

[Kalera et al., 2004] Kalera, M. K., Srihari, S., and Xu, A. (2004). Offline signature verification and identification using distance statistics. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(7):1339–1360.

[Kittler et al., 1998] Kittler, J., Hatef, M., Duin, R. P., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

[Mighell et al., 1989] Mighell, D. A., Wilkinson, T. S., and Goodman, J. W. (1989). Backpropagation and its application to handwritten signature verification. *Advances in Neural Information Processing Systems 1*, pages 340–347.

[Nemcek and Lin, 1974] Nemcek, W. F. and Lin, W. C. (1974). Experimental investigation of automatic signature verification. *IEEE Transactions on Systems, Man and Cybernetics*, 4:121–126.

[Oliveira et al., 2007] Oliveira, L. S., Justino, E. J. R., and Sabourin, R. (2007). Off-line signature verification using writer-independent approach. In *IJCNN*, pages 2539–2544.

[Ozgunduz et al., 2005] Ozgunduz, E., Senturk, T., and Karsligil, M. E. (2005). Off-line signature verification and recognition by support vector machine. In *13th European Signal Processing Conference (EUSIPCO 2005)*.

[Pekalska and Duin, 2002] Pekalska, E. and Duin, R. P. W. (2002). Dissimilarity representations allow for building good classifiers. *Pattern Recogn. Lett.*, 23(8):943–956.

[Plamondon and Lorette, 1989] Plamondon, R. and Lorette, G. (1989). Automatic signature verification and writer identification – the state of the art. *Pattern Recognition*, 22(2):107–131.

[Plamondon and Srihari, 2000] Plamondon, R. and Srihari, S. N. (2000). On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):63–84.

[Qi and Hunt, 1994] Qi, Y. and Hunt, B. R. (1994). Signature verification using global and grid features. *Pattern Recognition*, 27(12):1621–1629.

[Qi and Hunt, 1995] Qi, Y. and Hunt, B. R. (1995). A multiresolution approach to computer verification of handwritten signatures. *IEEE Transactions on Image Processing*, 4(6):870–874.

[Sabourin and Genest, 1994] Sabourin, R. and Genest, G. (1994). An extended-shadow-code based approach for off-line signature verification. i. evaluation of the bar mask definition. In *Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*, volume 2, pages 450–453 vol.2.

[Sabourin and Genest, 1995] Sabourin, R. and Genest, G. (1995). An extended-shadow-code based approach for off-line signature verification. ii. evaluation of several multi-classifier combination strategies. *ICDAR*, 01:197.

[Santini and Jain, 1999] Santini, S. and Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883.

[Santos, 2004] Santos, C. R. (2004). Análise de assinaturas manuscritas baseada nos princípios da grafoscopia. Master's thesis, Pontíficia Universidade Católica do Paraná.

[Sargur N. Srihari and Shah, 2007] Sargur N. Srihari, Chen Huang, H. S. and Shah, V. (2007). *Biometric and Forensic Aspects of Digital Document Processing*. Advances in Pattern Recognition: Digital Document Processing.

[Sproull, 1979] Sproull, R. F. (1979). *Principles of interactive computer graphics (2nd ed.)*. McGraw-Hill, Inc., New York, NY, USA.

[Srihari et al., 2004] Srihari, S. N., Xu, A., and Kalera, M. K. (2004). Learning strategies and classification methods for off-line signature verification. In *IWFHR 04, Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition*, pages 161–166, Washington, DC, USA. IEEE Computer Society.

[Ueda, 2003] Ueda, K. (2003). Investigation of off-line japanese signature verification using a pattern matching. In *ICDAR03*, pages 951–955.

[Vapnik, 1995] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.