

Rodrigo Tosta Peres

**Novas Técnicas de Classificação de Padrões baseadas
em Métodos Local-Global**

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Carlos Kubrusly

Rio de Janeiro, agosto de 2008

Rodrigo Tosta Peres

Novas Técnicas de Classificação de Padrões baseadas em Métodos Local-Global

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Dr. Carlos Kubrusly

Orientador

Departamento de Engenharia Elétrica/PUC-Rio

Dr. Raul Queiroz Feitosa

Departamento de Engenharia Elétrica/PUC-Rio

Dr. Marcelo da Cunha Medeiros

Departamento de Economia/PUC-Rio

Dr. Amit Bhaya

UFRJ

Dr. Marcello Luiz Rodrigues Campos

COPPE/UFRJ

Dr. Sérgio Lima Netto

COPPE/UFRJ

Dr. Alexandre P. Alves da Silva

UFRJ

Prof. José Eugênio Leal

Coordenador Setorial do Centro
Técnico Científico - PUC-Rio

Rio de Janeiro 11 de agosto de 2008

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Rodrigo Tosta Peres

Possui graduação em Licenciatura em Matemática pela Universidade Federal Fluminense (2000) e mestrado em Engenharia Elétrica pela Pontifícia Universidade Católica do Rio de Janeiro (2004). Tem experiência na área de Inteligência Computacional, com aplicações em Classificação de Padrões.

Ficha Catalográfica

Peres, Rodrigo Tosta

Novas técnicas de classificação de padrões baseadas em métodos local-global / Rodrigo Tosta Peres ; orientador: Carlos Kubrusly. – 2008.

81 f. : Il. ; 30 cm

Tese (Doutorado em Engenharia Elétrica)– Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Classificação de padrões. 3. Quantização vetorial. 4. Teoria da informação. 5. Máquinas de vetor de suporte. I. Kubrusly, Carlos. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Agradecimentos

A Deus, que sempre nos garante a certeza da vitória.

Ao meu orientador, Professor Carlos Eduardo Pedreira, por tudo que vem me ensinando ao longo desses anos de trabalho e pela parceria e amizade que permanecerão para sempre.

Ao professor Carlos Kubrusly, pela amizade e confiança de que este trabalho seria bem sucedido.

A PUC-Rio e ao CNPq, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Aos meus pais, Mário e Suely, e aos meus irmãos, Diego e Tiago, por terem me dado à base familiar que é tão importante na vida.

Ao professor Alexandre Pinto Alves da Silva, por todas as contribuições que foram muito importantes para o desenvolvimento desta tese.

Ao professor Moisés Henrique Szwarcman, pela amizade e por toda a ajuda administrativa na reta final do trabalho.

Aos amigos Hélio Pinto, André Sih e Francisco Carlos de Azevedo Pinto pelas contribuições dadas.

A banca desta tese por todo o trabalho de revisão e todas as sugestões que em muito contribuíram para este trabalho.

A todos os familiares, amigos e colegas que, direta ou indiretamente, contribuíram para que este trabalho pudesse ser desenvolvido.

Resumo

Peres, Rodrigo Tosta; Kubrusly, Carlos. **Novas Técnicas de Classificação de Padrões baseadas em Métodos Local-Global**. Rio de Janeiro, 2008. 81p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

O foco desta tese está direcionado a problemas de Classificação de Padrões. A proposta central é desenvolver e testar alguns novos algoritmos para ambientes supervisionados, utilizando um enfoque local-global. As principais contribuições são: (i) Desenvolvimento de método baseado em quantização vetorial com posterior classificação supervisionada local. O objetivo é resolver o problema de classificação estimando as probabilidades posteriores em regiões próximas à fronteira de decisão; (ii) Proposta do que denominamos ‘Zona de Risco Generalizada’, um método independente de modelo, para encontrar as observações vizinhas à fronteira de decisão; (iii) Proposta de método que denominamos ‘Quantizador Vetorial das Fronteiras de Decisão’, um método de classificação que utiliza protótipos, cujo objetivo é construir uma aproximação quantizada das regiões vizinhas à fronteira de decisão. Todos os métodos propostos foram testados em bancos de dados, alguns sintéticos e outros publicamente disponíveis.

Palavras-chave

Classificação de Padrões; Quantização Vetorial; Seleção de Observações; Teoria da Informação; Máquinas de Vetor de Suporte.

Abstract

Peres, Rodrigo Tosta; Kubrusly, Carlos. **New Techniques of Pattern Classification based on Local-Global Methods**. Rio de Janeiro, 2008. 81p. PhD.Thesis - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This thesis is focused on Pattern Classification problems. The objective is to develop and test new supervised algorithms with a local-global approach. The main contributions are: (i) A method based on vector quantization with posterior supervised local classification. The classification problem is solved by the estimation of the posterior probabilities near the decision boundary; (ii) Propose of what we call ‘Zona de Risco Generalizada’, an independent model method to find observations near the decision boundary; (iii) Propose of what we call ‘Quantizador Vetorial das Fronteiras de Decisão’, a classification method based on prototypes that build a quantized approximation of the decision boundary. All methods were tested in synthetics or real datasets.

Keywords

Pattern Classification; Vector Quantization; Observations Selection; Information Theory; Support Vector Machines.

Sumário

1	Introdução	14
2	Metodologia e Dados	19
2.1.	Classificador por Quantização Vetorial (CQV)	21
2.2.	A Zona de Risco Generalizada: Seleção de Observações para melhorar a Classificação	26
2.2.1.	Uma Aplicação da ZRG a Seleção de Observações	32
2.3.	Quantizador Vetorial das Fronteiras de Decisão (QVFD)	34
2.4.	Sobre os Dados	38
2.4.1.	Classes divididas através de função cosseno	38
2.4.2.	Quadrados Sobrepostos	39
2.4.3.	Círculos Concêntricos	39
2.4.4.	Waveform	40
2.4.5.	Reconhecimento da Letra B	40
2.4.6.	Statlog – Landsat Satellite	41
2.4.7.	Diagnóstico de Doenças Cardíacas	41
2.4.8.	Classificação de Punção Aspirativa por Agulha Fina de Nódulo Mamário para Diagnóstico	41
2.4.9.	Ionosfera	42
2.4.10.	Câncer de Pulmão	42
3	Resultados e Discussão	43
3.1.	Resultados Referentes à Classificação	43
3.1.1.	Classificador por Quantização Vetorial (CQV) e Quantizador Vetorial das Fronteiras de Decisão (QVFD)	43
3.1.2.	Zona de Risco Generalizada	49
3.2.	Resultados referentes à Seleção de Observações	60
3.2.1.	Uma Aplicação do Classificador por Quantização Vetorial (CQV) à Seleção de Observações	60
3.2.2.	Resultados da Aplicação de Zona de Risco Generalizada (ZRG)	

em Seleção de Observações	67
4 Conclusão	72
5 Referências	74
6 Apêndice A – Notação	78
7 Apêndice B – Quantização Vetorial	79

Lista de figuras

Figura 1: Problema de classificação não separável linearmente com uma solução global.	20
Figura 2: Problema de classificação não separável linearmente com 8 sub-soluções triviais ou separáveis linearmente.	20
Figura 3: Célula S_k dividida em duas classes com suas médias m_1 e m_2 .	22
Figura 4: A observação x em S_k classificada de acordo com sua distância às médias.	22
Figura 5: Uma observação classificada erroneamente em virtude do critério utilizado.	23
Figura 6: Duas classes divididas pela função cosseno. Observações em vermelho são aquelas cujo rótulo está invertido artificialmente.	30
Figura 7: ZRG com regiões ruidosas. $\Sigma = 0.1$.	31
Figura 8: ZRG depois do procedimento de limpeza. $\Sigma = 0.1$.	31
Figura 9: Resultado do QVFD para o banco de dados apresentado na seção 2.4.1.	36
Figura 10: Classes C_1 e C_2 divididas através de função cosseno.	38
Figura 11: Classes C_1 e C_2 – Quadrados Sobrepostos.	39
Figura 12: Classes C_1 e C_2 – Círculos Concêntricos.	40
Figura 13: ZRG para o banco de dados apresentado na seção 2.4.1. Σ obtido pela validação cruzada.	56
Figura 14: ZRG para o banco de dados apresentado na seção 2.4.1. Σ obtido pela equação 4.14 de [41].	57
Figura 15: ZRG para o banco de dados apresentado na seção 2.4.1. Σ obtido pela média entre os sigmas de validação cruzada e equação 4.14 de [41].	57
Figura 16: ZRG para o banco de dados apresentado na seção 2.4.1. $\Sigma = 0.1$.	58
Figura 17: ZRG para o banco de dados apresentado na seção 2.4.3. Σ obtido pela validação cruzada.	58
Figura 18: ZRG para o banco de dados apresentado na seção 2.4.3. Σ obtido pela equação 4.14 de [41].	59

Figura 19: ZRG para o banco de dados apresentado na seção 2.4.3. Sigma obtido pela média entre os sigmas de validação cruzada e equação 4.14 de [41].	59
Figura 20: ZRG para o banco de dados apresentado na seção 2.4.3. Sigma 0.1.	60
Figura 21: Classes C_1 e C_2 divididas através de função cosseno. Para cada classe, 20 observações (círculos vermelhos) tiveram seu rótulo invertido.	61
Figura 22: Classes C_1 e C_2 – Círculos Concêntricos com 5 e 20 observações com rótulo invertido (círculos vermelhos).	63
Figura 23: Classes C_1 e C_2 – Quadrados Sobrepostos. Para cada classe, 20 observações (círculos vermelhos) tiveram seu rótulo invertido.	65
Figura 24: Classes C_1 e C_2 – Quadrados Sobrepostos. Observações mantidas pelo método CQV Aplicado a Seleção de Observações.	65
Figura 25: Classes C_1 e C_2 – Quadrados Sobrepostos. Observações mantidas pelo método ZRG Aplicado a Seleção de Observações.	69

Lista de tabelas

Tabela 1: Desempenho de classificação fora-da-amostra (Acerto) e desvio padrão (dp) para os algoritmos CQV e QVFD. Parênteses indicam dentro-da-amostra.	46
Tabela 2: Tempo computacional médio em segundos.	46
Tabela 3: Custo computacional em segundos para procedimento de limpeza, obtenção dos sigmas e ZRG's.	51
Tabela 4: Taxa de acerto fora-da-amostra para LVQ (parênteses indicam dentro-da-amostra).	52
Tabela 5: Taxa de acerto fora-da-amostra para LVQ (Zona de Risco).	52
Tabela 6: Taxa de acerto fora-da-amostra para RN (parênteses indicam dentro-da-amostra).	53
Tabela 7: Taxa de acerto fora-da-amostra para SVM (parênteses indicam dentro-da-amostra).	53
Tabela 8: Taxa de acerto fora-da-amostra para K-NN (parênteses indicam dentro-da-amostra).	54
Tabela 9: Experimentos com IB3 e DROP3.	54
Tabela 10: AUC para RN.	55
Tabela 11: Experimento 19: Reconhecimento de rótulos invertidos.	62
Tabela 12: Experimento 19: Reconhecimento de rótulos invertidos por DROP3 e IB3.	62
Tabela 13: Experimento 20: Reconhecimento de rótulos invertidos.	63
Tabela 14: Experimento 20: Reconhecimento de rótulos invertidos por DROP3 e IB3.	64
Tabela 15: Experimento 21: Reconhecimento de rótulos invertidos.	66
Tabela 16: Experimento 21: Reconhecimento de rótulos invertidos por DROP3 e IB3.	66
Tabela 17: Observações classificadas em típicas (T), de risco (R) ou com inversão de rótulo (Inv). Entre parênteses, a porcentagem que representam em relação ao total de observações.	67
Tabela 18: Observações classificadas como Inv que de fato tiveram o	

rótulo invertido e observações classificadas como Inv que não tiveram o rótulo invertido.	68
Tabela 19: Observações classificadas em típicas (T), de risco (R) ou com inversão de rótulo (Inv). Entre parênteses, a porcentagem que representam em relação ao total de observações.	68
Tabela 20: Observações classificadas como Inv que de fato tiveram o rótulo invertido e observações classificadas como Inv que não tiveram o rótulo invertido.	69
Tabela 21: Observações classificadas em típicas (T), de risco (R) ou com inversão de rótulo (Inv). Entre parênteses, a porcentagem que representam em relação ao total de observações.	70
Tabela 22: Observações classificadas em típicas (T), de risco (R) ou com inversão de rótulo (Inv). Entre parênteses, a porcentagem que representam em relação ao total de observações.	70
Tabela 23: Observações classificadas como Inv que de fato tiveram o rótulo invertido e observações classificadas como Inv que não tiveram o rótulo invertido.	70

Ora, àquele que é poderoso para fazer tudo muito mais abundantemente além daquilo que pedimos ou pensamos, segundo o poder que em nós opera, a esse glória na igreja, por Jesus Cristo, em todas as gerações, para todo o sempre.

Amém!

(Efésios 3:20-21)