PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Rodrigo Tosta Peres

Seleção de Dados em LVQ

Dissertação de Mestrado

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

Orientador: Carlos Eduardo Pedreira

Rio de Janeiro, agosto de 2004

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Rodrigo Tosta Peres

Seleção de Dados em LVQ

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

Carlos Eduardo Pedreira
Orientador
PUC-Rio

Álvaro Veiga
PUC-Rio

Alexandre P. Alves da Silva
UFRJ

Felipe França
UFRJ

José Eugênio Leal
Coordenador(a) Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 12 de agosto de 2004

Rodrigo Tosta Peres

Graduado em Licenciatura em Matemática pela UFF (Universidade Federal Fluminense), em 2000.

Ficha Catalográfica

Peres, Rodrigo Tosta

Seleção de dados em LVQ / Rodrigo Tosta Peres; orientador: Carlos Eduardo Pedreira. – Rio de Janeiro: PUC, Departamento de Engenharia Elétrica, 2004.

68 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Inclui referências bibliográficas.

1. Engenharia elétrica – Teses. 2. Classificação de padrões. 3. Seleção de dados. 4. LVQ. 5. Aprendizado por quantização vetorial. 6. Active learning. 7. Aprendizado exaustivo. 8. Redes neurais. I. Pedreira, Carlos Eduardo. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Aos meus pais, Mário e Suely,
pelo amor incondicional durante todos esses anos.

Agradecimentos

A Deus, que está comigo durante todo o tempo.

Ao meu orientador, Professor Carlos Eduardo Pedreira, pelo apoio em todas as horas, pelo espírito científico que vem me ajudando a desenvolver, pela confiança e, acima de tudo, pela amizade.

A PUC-Rio, pelos auxílios concedidos, sem os quais esse trabalho não poderia ter sido realizado.

Aos meus pais e aos meus irmãos, Diego e Tiago, por todo o amor, carinho e paciência, além de terem me dado o suporte para que eu pudesse chegar até aqui.

A Professora Ana Pavani, por tudo que me ensinou durante os anos de trabalho, pela amizade e respeito que ficarão para sempre.

Ao Professor Álvaro Veiga pela força e pelo conselho de ir em frente mesmo nas horas difíceis.

Ao Professor Yaser Abu-Mostafa e a sua equipe pelo apoio técnico durante o desenvolvimento do trabalho, especialmente ao Hsuan-Tien Lin pela troca de idéias sempre interessante.

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

Resumo

Peres, Rodrigo Tosta; Pedreira, Carlos Eduardo. **Seleção de Dados em LVQ.** Rio de Janeiro, 2004. 68p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Nesta dissertação, propomos uma metodologia para seleção de dados em modelos de Aprendizado por Quantização Vetorial, referenciado amplamente na literatura pela sigla em inglês LVQ. Treinar um modelo (ajuste dentro-da-amostra) com um subconjunto selecionado a partir do conjunto de dados disponíveis para o aprendizado pode trazer grandes benefícios no resultado de generalização (fora-da-amostra). Neste sentido, é muito importante realizar uma busca para selecionar dados que, além de serem representativos de suas distribuições originais, não sejam ruído (no sentido definido ao longo desta dissertação). O método proposto procura encontrar os pontos relevantes do conjunto de entrada, tendo como base a correlação do erro de cada ponto com o erro do restante da distribuição. Procura-se, em geral, eliminar considerável parte do ruído mantendo os pontos que são relevantes para o ajuste do modelo (aprendizado). Assim, especificamente em LVQ, a atualização dos protótipos durante o aprendizado é realizada com um subconjunto do conjunto de treinamento originalmente disponível. Experimentos numéricos foram realizados com dados simulados e reais, e os resultados obtidos foram muito interessantes, mostrando claramente a potencialidade do método proposto.

Palavras-chave

Classificação de Padrões; Seleção de Dados; LVQ; Aprendizado por Quantização Vetorial; Active Learning; Aprendizado Exaustivo; Redes Neurais.

Abstract

Peres, Rodrigo Tosta; Pedreira, Carlos Eduardo. **Data Selection for LVQ.** Rio de Janeiro, 2004. 68p. MSc. Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

In this dissertation, we consider a methodology for selection of data in models of Learning Vector Quantization (LVQ). The generalization can be improved by using a subgroup selected from the available data set. We search the original distribution to select relevant data that aren't noise. The search aims at relevant points in the training set based on the correlation between the error of each point and the average of error of the remaining data. In general, it is desired to eliminate a considerable part of the noise, keeping the points that are relevant for the learning model. Thus, specifically in LVQ, the method updates the prototypes with a subgroup of the originally available training set. Numerical experiments have been done with simulated and real data. The results were very interesting and clearly indicated the potential of the method.

Keywords

Pattern Classification; Data Selection; LVQ; Learning Vector Quantization; Active Learning; Exaustive Learning; Neural Networks.

Sumário

Lista de figuras

Lista de tabelas

A sabedoria é suprema; portanto, adquire a sabedoria.
Sim, com tudo o que possuis adquire o entendimento.
(Provérbios 4:7)