

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO



Raúl Pierre Rentería

**Algoritmos para regressão por mínimos
quadrados parciais**

Tese de Doutorado

Tese apresentada ao Programa de Pós-graduação em
Ciência da Computação do Departamento de Informática
da PUC-Rio como parte dos requisitos parciais para
obtenção do título de Doutor em Ciência da Computação

Orientador: Prof. Ruy Luiz Milidiú

Rio de Janeiro
Março de 2003



Raúl Pierre Rentería

**Algoritmos para regressão por mínimos
quadrados parciais**

Tese apresentada ao Programa de Pós-graduação em
Ciência da Computação do Departamento de Informática
do Centro Técnico Científico da PUC-Rio como parte dos
requisitos parciais para obtenção do título de Doutor em
Ciência da Computação. Aprovada pela Comissão Exami-
nadora abaixo assinada.

Prof. Ruy Luiz Milidiú

Orientador

Departamento de Informática — PUC-Rio

Prof. Carlos José Pereira de Lucena

PUC-Rio

Prof. Carlos Eduardo Pedreira

PUC-Rio

Prof. Valmir Carneiro Barbosa

UFRJ

Prof. Fernando Antônio de Carvalho Gomes

UFC

Prof. Luiz Pereira Calôba

UFRJ

Prof. Eduardo Sany Laber

PUC-Rio

Prof. Ney Augusto Dumont

Coordenador Setorial do Centro Técnico Científico —

PUC-Rio

Rio de Janeiro, 19 de Março de 2003

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

Raúl Pierre Rentería

Graduou-se em Engenharia de Computação na Pontifícia Universidade Católica do Rio de Janeiro em 1996.

Ficha Catalográfica

Rentería, Raúl Pierre

Algoritmos para regressão por mínimos quadrados parciais / Raúl Pierre Rentería; orientador: Ruy Luiz Milidiú. — Rio de Janeiro : PUC-Rio, Departamento de Informática, 2003.

80 f. : il. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Teses. 2. PLS
3. Mínimos quadrados parciais. 4. Paralelismo 5. Aproximação. 6. Regressão não linear. 7. Funções de núcleo 8. Multi-Kernel. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

À minha família pelo apoio incondicional.

Ao Christian e ao Sacha pelas revisões e sugestões.

Ao Eduardo Laber, não somente pelos conselhos amigos mas também pelo exemplo de dedicação acadêmica.

Ao meu orientador Professor Ruy Milidiú pelo estímulo e parceria para este trabalho.

Aos amigos do LEARN, em especial Artur e Fred, pelas inúmeras conversas e comentários a respeito de minha tese.

Aos professores que participarem da Comissão examinadora.

Ao Brumado, velho amigo de Santa Teresa pelas longas conversas que inspiraram meu trabalho.

A todos os professores e funcionários do Departamento pelos ensinamentos e pela ajuda.

Ao CNPQ e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Resumo

Rentería, Raúl Pierre; Milidiú, Ruy Luiz. **Algoritmos para regressão por mínimos quadrados parciais**. Rio de Janeiro, 2003. 80p. Tese de Doutorado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Muitos problemas da área de aprendizagem automática tem por objetivo modelar a complexa relação existente num sistema, entre variáveis de entrada X e de saída Y na ausência de um modelo teórico. A regressão por mínimos quadrados parciais PLS (Partial Least Squares) constitui um método linear para resolução deste tipo de problema, voltado para o caso de um grande número de variáveis de entrada quando comparado com o número de amostras.

Nesta tese, apresentamos uma variante do algoritmo clássico PLS para o tratamento de grandes conjuntos de dados, mantendo um bom poder preditivo. Dentre os principais resultados destacamos uma versão paralela PPLS (Parallel PLS) exata para o caso de apenas uma variável de saída e uma versão rápida e aproximada DPLS (Direct PLS) para o caso de mais de uma variável de saída.

Por outro lado, apresentamos também variantes para o aumento da qualidade de predição graças à uma formulação não linear. São elas o LPLS (Lifted PLS), algoritmo para o caso de apenas uma variável de saída, baseado na teoria de funções de núcleo (kernel functions), uma formulação kernel para o DPLS e um algoritmo multi-kernel MKPLS capaz de uma modelagem mais compacta e maior poder preditivo, graças ao uso de vários núcleos na geração do modelo.

Palavras-chave

PLS; Regressão por mínimos quadrados parciais; DPLS; Paralelismo; PPLS; MKPLS; Regressão não linear; Funções de núcleo; Multi-kernel.

Abstract

Rentería, Raúl Pierre; Milidiú, Ruy Luiz. **Algorithms for Partial Least Squares regression**. Rio de Janeiro, 2003. 80p. PhD. Thesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The purpose of many problems in the machine learning field is to model the complex relationship in a system between the input X and output Y variables when no theoretical model is available. The Partial Least Squares (PLS) is one linear method for this kind of problem, for the case of many input variables when compared to the number of samples.

In this thesis we present versions of the classical PLS algorithm designed for large data sets while keeping a good predictive power. Among the main results we highlight PPLS (Parallel PLS), a parallel version for the case of only one output variable, and DPLS (Direct PLS), a fast and approximate version, for the case of more than one output variable.

On the other hand, we also present some variants of the regression algorithm that can enhance the predictive quality based on a non-linear formulation. We introduce LPLS (Lifted PLS), for the case of only one dependent variable based on the theory of kernel functions, KDPLS, a non-linear formulation for DPLS, and MKPLS, a multi-kernel algorithm that can result in a more compact model and a better prediction quality, thanks to the use of several kernels for the model building.

Keywords

PLS; Partial Least Squares; DPLS; Paralelism; PPLS; MKPLS; Non linear regression; Kernel functions; Multi-kernel.

Sumário

1	Introdução	10
2	Regressão por mínimos quadrados parciais	12
2.1	Modelo	12
2.2	Algoritmo	14
2.3	Propriedades	16
2.4	Seleção de fatores	20
3	Ambiente experimental	22
3.1	Conjunto NIR	22
3.2	Conjunto Santa Fé	25
4	Algoritmos para grandes conjuntos de dados	27
4.1	PPLS	27
4.2	DPLS	32
5	Algoritmos baseados em funções de Núcleo	40
5.1	Funções de Núcleo	40
5.2	LPLS	49
5.3	KDPLS	55
5.4	MKPLS - PLS Multi Núcleos	61
6	Conclusão	72
6.1	Principais resultados	72
6.2	Trabalhos Futuros	74

Lista de Figuras

2.1	Algoritmo PLS2.	15
2.2	Algoritmo PLS1.	16
2.3	Predição PLS.	17
2.4	Gráfico do PRESS em função do número de fatores.	21
3.1	Fragmento da série D utilizada nos experimentos.	26
4.1	Cálculo de $\{w, b, p\}$ para o nó i .	29
4.2	Cálculo de $\{w, b, p\}$ para o nó 0.	29
4.3	Tempo de execução do PPLS.	30
4.4	Speedup para o PPLS.	31
4.5	Efficiency para o PPLS.	31
4.6	PRESS dos modelos PLS e DPLS para o conjunto Wheat.	36
4.7	PRESS dos modelos PLS e DPLS para o conjunto Wheat, para os fatores escolhidos.	37
4.8	PRESS dos modelos PLS e DPLS para o conjunto Metal ions.	38
4.9	PRESS dos modelos PLS e DPLS para o conjunto Combustible.	39
5.1	Exemplo de mapeamento desejado.	41
5.2	Reformulação KBPLS para o algoritmo NIPALS.	46
5.3	Algoritmo LPLS.	52
5.4	Predição LPLS.	53
5.5	Comparação do LPLS com o PLS.	55
5.6	PRESS dos modelos KBPLS e KDPLS para o conjunto Light Gas Oil.	59
5.7	PRESS dos modelos KBPLS e KDPLS para o conjunto Light Gas Oil.	60
5.8	PRESS dos modelos KBPLS e KDPLS para o conjunto Polymer.	60
5.9	PRESS dos modelos KBPLS e KDPLS para o conjunto Wheat.	61
5.10	Algoritmo MKPLS para desconto no treinamento.	62
5.11	Algoritmo MKPLS para desconto na predição.	63
5.12	Principais passos do MKPLS.	64
5.13	Curva PRESS de PLS, LPLS e MKPLS para o conjunto Wheat.	65
5.14	Curva PRESS de PLS, LPLS e MKPLS para o conjunto Meat.	66
5.15	Curva PRESS de PLS, LPLS e MKPLS para o conjunto Com- bustible.	68
5.16	Curva PRESS de PLS, LPLS e MKPLS para o conjunto Light gas oil.	69
5.17	Curva PRESS de PLS, LPLS e MKPLS para o conjunto Corn.	70

Lista de Tabelas

3.1	Conjuntos de dados	23
4.1	Desempenho do PPLS	30
4.2	PRESS e número de fatores escolhidos para o PLS e o DPLS	37
5.1	Comparação entre KBPLS e KDPLS com núcleo polinomial	58
5.2	Comparação entre KBPLS e KDPLS com núcleo gaussiano	58
5.3	Comparação do PRESS do MKPLS com PLS e LPLS para os primeiros fatores.	67
5.4	Comparação do PRESS do MKPLS com PLS e LPLS para todos os fatores.	67