1
Introdução

Muitos problemas da área de aprendizagem automática têm 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) desenvolvida por Wold et al. [56, 57], constitui um método linear para resolução deste tipo de problemas, voltado para o caso de um grande número de variáveis de entrada quando comparado com o número de amostras. Com relação à regressão linear clássica OLSR [1] que emprega todas as variáveis de entrada simultaneamente, a regressão PLS fornece resultados mais robustos por realizar a regressão de forma incremental apenas num conjunto das variáveis de entrada.

O uso da regressão PLS tem se desenvolvido nas últimas duas décadas, sendo que nos últimos anos tem recebido atenção específica em congressos especializados neste tema [43, 44]. O sucesso alcançado na sua área de desenvolvimento inicial, a quimiometria [22, 23], induziu sua aplicação em outras áreas tais como monitoração de processos, marketing e processamento de imagens [52].

Nesta tese, são apresentadas variantes do algoritmo clássico PLS com o objetivo de fornecer métodos capazes de lidar com aplicações diversas e atuais, permitindo que seja alcançado o mesmo sucesso encontrado na área de processos químicos. Para tanto, dois pontos são abordados.

Preocupado com o tamanho crescente dos dados nas aplicações atuais [27], o primeiro ponto tratado refere-se à escalabilidade dos algoritmos. Neste sentido, são apresentadas alternativas para o tratamento de grandes conjuntos de dados, mantendo um bom poder preditivo. Para o caso de apenas uma variável de saída, conhecido como PLS1, é fornecido PPLS[2] [37, 41, 40], uma versão paralela distribuída de fácil implementação. Para aplicações com mais de uma variável de saída, também conhecidas como

[1] Ordinary Least Squares Regression
[2] Parallel PLS

PLS2, é apresentado DPLS[3] [37, 41, 36], uma versão aproximada para o algoritmo de regressão PLS2. Os resultados obtidos mostram que a aproximação fornece ganho de desempenho computacional expressivo, mantendo uma qualidade de predição competitiva quando comparada com a do algoritmo exato.

O segundo ponto desenvolvido trata de formulações não-lineares baseadas em funções de núcleo[4] [6], visando o aumento da qualidade de predição. Após uma estudo da aplicação de funções de núcleo para a incorporação de não-linearidade em algoritmos, é apresentado o LPLS[5], versão PLS1 mais eficiente quando comparada com implementações existentes [46]. De forma semelhante à realizada para a versão linear, é mostrado KDPLS, uma versão *kernel* para o DPLS apresentando as mesmas vantagens. Estendendo LPLS e KDPLS como outras versões não-lineares baseadas em um núcleo [46], é apresentado o algoritmo MKPLS[6] [38, 39], capaz de uma modelagem mais compacta e de maior poder preditivo graças ao uso de vários núcleos na geração do modelo. Todos os algoritmos não-lineares desenvolvidos se beneficiam de uma formulação simples conseguida com funções de núcleo.

A tese está organizada da forma descrita a seguir. No capítulo 2, é apresentado o algoritmo PLS nas duas formas encontradas na literatura [52]: PLS2 quando existem mais de uma variável de saída e PLS1 para o caso de apenas uma. Além disto são fornecidas as principais propriedades usadas neste trabalho. Após descrever, no capítulo 3, os conjuntos de dados usados para a avaliação dos vários algoritmos desenvolvidos, são apresentadas no capítulo 4 as variantes PPLS e DPLS para o tratamento de grandes conjuntos de dados. Finalmente, no capítulo 5 são apresentadas as variantes não-lineares baseadas em funções de núcleo: LPLS, KDPLS e MKPLS.

[3] Direct PLS
[4] Kernel functions
[5] Lifted PLS
[6] Multi-Kernel PLS