

Referência do projecto: PTDC/EEA-TEL/72572/2006

Título do projecto: *Desenvolvimento e Aprendizagem de Núcleos para Texto e Imagens / Development and Learning of Kernels for Text and Images (DeLKeTI)*

Data de início do projecto: 01 / Dezembro / 2007

1. Introdução

Os objectivos centrais do projecto situavam-se na área do desenvolvimento e aprendizagem de *kernels* (núcleos), os quais são componentes fundamentais de muitas das modernas técnicas de "machine learning", nomeadamente das famosas "support vector machines". A utilização de *kernels* provocou uma verdadeira revolução quer na teoria quer na prática da aprendizagem automática (*machine learning*), nomeadamente por permitir criar versões não lineares de todos os algoritmos clássicos lineares. Recentemente, estes métodos foram generalizados com grande sucesso para dados não vectoriais (por exemplo, sequências de símbolos, árvores, grafos, imagens). As duas vertentes/objectivos principais do projecto eram: (a) desenvolvimento e aplicação de *kernels* baseados em teoria da informação e compressão, nomeadamente para classificação de sequências (por exemplo, textos). (b) Desenvolvimento de técnicas capazes de aprender *kernels* directamente a partir de dados observados.

2. *Kernels* Baseados em Teoria da Informação Não Extensiva

Nesta vertente do projecto conduziu-se investigação quer de carácter teórico quer de índole mais aplicada. Na frente de trabalho mais teórica, desenvolveu-se uma nova classe de *kernels* baseados em teoria da informação não extensiva (também referida como teoria da informação de Tsallis), tendo sido mostrado como obter *kernels* definidos-positivos com base nessa abordagem. Os *kernels* obtidos aplicam-se a medidas (normalizadas, ou seja, distribuições de probabilidade, ou não normalizadas) e estendem consideravelmente uma classe de *kernels* anteriormente existente baseada na divergência de Kullback-Leibler (isto é, contendo esta classe como caso particular). Os resultados desta frente de investigação culminaram na publicação de um extenso artigo (de cerca de 40 páginas) na mais prestigiada revista da área de "machine learning":

- A. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo, “Nonextensive information theoretic *kernels* on measures”, *Journal of Machine Learning Research*, vol. 10, pp. 935 – 975, 2009.

No seguimento dos desenvolvimento deste *kernels* baseados em teoria da informação não extensiva, estes foram de seguida aplicados em vários problemas práticos de reconhecimento automático de padrões; estas aplicações foram descritas numa série de artigos de conferência, em colaboração e co-autoria com o grupo do Prof. Vittorio Murino, da Universidade de Verona (Itália):

- M. Figueiredo, P. Aguiar, A. Martins, V. Murino, M. Bicego, “Information theoretical kernels for generative embeddings based on hidden Markov models”, *Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition - S+SSPR'2010*, Cesme, Izmir, Turkey, 2010.
- M. Bicego, A. Perina, V. Murino, A. Martins, P. Aguiar, M. Figueiredo, “Combining free energy score spaces with information theoretic kernels: application to scene classification”, *IEEE International Conference on Image Processing – ICIP'2010*, Hong Kong, 2010.
- M. Bicego, A. Martins, V. Murino, P. Aguiar, and M. Figueiredo, “2D shape recognition using information theoretic kernels”, *IAPR International Conference on Pattern Recognition – ICPR'2010*, Istanbul, Turkey, 2010.

Dada a importância e profundidade dos resultados reportados no artigo publicado no JMLR acima referido, bem como a competitividade dos resultados aplicados reportados nas outras três publicações, pode argumentar-se que, mesmo que tivesse sido esta a única vertente do projecto, este poderia mesmo assim ser considerado como plenamente bem sucedido.

3. *Kernels* Baseados em Técnicas de Compressão.

Os *kernels* baseados em teoria da informação não extensiva, descritos na secção anterior, exigem a estimação de modelos probabilísticos para os objectos em presença (por exemplo, imagens, formas, textos), o que, em algumas circunstâncias, os pode tornar sensíveis à escolha da classe de modelos. Uma alternativa explorada no projecto consiste em aproximar as medidas (de semelhança) de teoria da informação nas quais se baseiam os *kernels* utilizando técnicas de compressão universal (do tipo Lempel-Ziv), isto é, que não se suportam em conhecimento prévio dos modelos probabilísticos das fontes.

Nesta vertente do trabalho, explorou-se a aplicabilidade desta abordagem a dois tipos de problemas diferentes (com tipos de dados diferentes): a classificação de textos (nomeadamente a atribuição de autoria) e a biometria. O trabalho levado a cabo em classificação de textos conduziu ao seguinte artigo:

- D. Pereira-Coutinho and M. Figueiredo, “Information-theoretic text classification: method evaluation”, *8th International Workshop on Pattern Recognition in Information Systems – PRIS'2008*, Barcelona, Spain, 2008.

Em termos da aplicação das medidas de semelhança baseadas em compressão universal a problemas de biometria, abordaram-se a autenticação e a identificação de pessoas baseadas em sinais de electrocardiografia (ECG). Foram estudadas várias alternativas para transformar o sinal de base numa sequência simbólica adequada às técnicas de compressão universal, nomeadamente a quantização uniforme e a quantização de Lloy-Max, baseada no algoritmo clássico de Linde-Buzo-Gray. Esta vertente do trabalho conduziu às três seguintes publicações:

- D. Pereira Coutinho, A. Fred, and M. Figueiredo, “One-lead ECG-based personal identification using Ziv-Merhav cross parsing”, *IAPR International Conference on Pattern Recognition – ICPR’2010*, Istanbul, Turkey, 2010.
- D. Pereira Coutinho, A. Fred, and M. Figueiredo, “Personal identification and authentication based on one-lead ECG by using Ziv-Merhav cross parsing”, *10th International Workshop on Pattern Recognition in Information Systems - PRIS 2010*, Funchal, Portugal, 2010.
- D. Pereira Coutinho, A. Fred, M. Figueiredo, "ECG-based continuous authentication using adaptive string matching", *International Conference on Bio-inspired Systems and Signal Processing - BIOSIGNALS'2011*, Rome, Italy, 2011.

Ainda nesta vertente do trabalho, foi submetido, estando em fase de revisão, o seguinte artigo:

- D. Pereira Coutinho, A. Fred, M. Figueiredo, "Fiducial and Non-fiducial Approaches to ECG-based Biometric Systems", *Pattern Recognition Letters*, 2011 (submitted).

Finalmente, e com o objectivo de obter versões rápidas dos algoritmos de compressão nos quais se baseam as medidas de semelhança em causa, exploraram-se técnicas baseadas em "suffix arrays"; este trabalho resultou na publicação de dois artigos (um dos quais em revista e o outro na prestigiada conferência *DCC - Data Compression Conference*):

- A. Ferreira, A. Oliveira, M. Figueiredo, "On the suitability of suffix arrays for Lempel-Ziv data compression", *Communications in Computer and Information Science*, vol. 48, pp. 267 - 280, 2009.
- A. Ferreira, A. Oliveira, M. Figueiredo, “On the use of suffix arrays for memory-efficient Lempel-Ziv data compression”, *The Data Compression Conference – DCC’2009*, Snowbird, UT, USA, 2009.

4. Aprendizagem de *Kernels*

A terceira, e final, vertente do projecto focava o problema da aprendizagem do *kernel* a partir de dados observados. O desempenho dos métodos baseados em *kernels* (nomeadamente classificadores) depende crucialmente da capacidade destes de exprimirem as medidas de semelhanças relevantes para o problema em causa;

este facto implica a necessidade de aperfeiçoar os *kernels*, quer por intervenção humana com base em conhecimento acerca dos dados em presença, quer usando métodos iterativos de refinamento do *kernel* (baseados em medidas do tipo "cross-validation"), que são extremamente morosos e computacionalmente pesados. Uma alternativa consiste em considerar uma combinação ponderada de diferentes *kernels* (num paradigma conhecido como "multiple *kernel* learning" - MKL), integrando a aprendizagem dos respectivos pesos no algoritmo de aprendizagem subjacente (por exemplo, aprendizagem de "support vector machines" - SVM - ou de "conditional random fields" - CRF).

A abordagem MKL, se bem que resolvendo parcialmente o problema da aprendizagem dos *kernels*, é ainda demasiado pesada computacionalmente para poder ser usada em problemas estruturados de larga escala, como o são a maioria dos problemas de processamento de linguagem/texto natural (por exemplo, "dependency parsing"). No âmbito do projecto, foi desenvolvido um novo algoritmo de MKL, combinando passos de sub-gradiente com passos proximais, cujas principais características são as seguintes:

- funciona de modo incremental ("on-line");
- é simples, flexível e compatível com uma vasta gama de regularizadores (esparsos e não esparsos);
- suporta regularizadores compostos (mesmo que com soberposições);
- é adequado para problemas de larga escala;
- adequa-se a problemas cuja estrutura varia no tempo.

Este trabalho foi descrito em detalhe em dois artigos:

- A. Martins, N. Smith, E. Xing, P. Aguiar, M. Figueiredo, "Online learning of structured predictors with multiple kernels", *Fourteenth International Conference on Artificial Intelligence and Statistics - AISTATS'2011*, Fort Lauderdale, FL, USA, 2011.
- A. Martins, N. Smith, E. Xing, P. Aguiar, M. Figueiredo, "Online MKL for Structured Prediction", *NIPS 2010 Workshop on New Directions in Multiple Kernel Learning*, Whistler, Canada, 2010.

Finalmente, uma versão de revista encontra-se em fase de revisão:

- A. Martins, N. Smith, E. Xing, P. Aguiar, M. Figueiredo, "Online Multiple Kernel Learning for Structured Prediction", *Journal of Machine Learning Research*, 2011 (submitted).

4. Formação Avançada

A orientação de alunos foi também parte integrante do projecto. O Prof. Mário Figueiredo é orientador dos estudantes de doutoramento André Martins e José David Pereira Coutinho, que realizaram trabalho de investigação no âmbito deste projecto (como se pode constatar pela autoria das publicações acima referidas).

Ambos os estudantes se encontram em fase de escrita das respectivas teses de doutoramento, com entrega planeada para o início de 2012.

5. Conclusões

Por tudo o que foi apresentado acima pensamos poder afirmar-se que o projecto foi plenamente bem sucedido. Recapitulando:

- a) Foram desenvolvidos novos *kernels* baseados em teoria da informação, nomeadamente em teoria da informação não extensiva e em algoritmos de compressão.
- b) Os *kernels* referidos na alínea anterior foram aplicados com sucesso a vários tipos de problemas: classificação de texto, de imagens, de formas e de sinais (nomeadamente, biométricos).
- c) Foi desenvolvido um novo algoritmo de aprendizagem de kernels a partir dos dados observados, baseado no paradigma MKL ("multiple kernel learning"), com aplicação (bem sucedida) a vários problemas de processamento e análise de linguagem/texto natural.
- d) Foram publicados artigos descrevendo todos os avanços conseguidos no âmbito do projecto em várias conferências de elevado prestígio e em vários artigos em revistas internacionais.
- e) Os dois investigadores estudantes de doutoramento que trabalharem no âmbito do projecto estão em fase final de escrita das respectivas teses de doutoramento.