Research summary

My research deals with machine learning methods, in particular multi-task learning and regularization-based methods, and their application to biological problems.

1 Pairwise learning for interaction prediction

In vaccine design, immunologists are interested in having accurate predictions of which peptides bind to MHC molecules. This is crucial to discover which peptides of a pathogen can trigger an immunological response and therefore give protection against the given pathogen. Different MHC alleles bind different peptides.

In drug discovery, biologists try to find small molecules which interact with given therapeutical targets such as enzymes or GPCRs. The goal is to use these molecules as drugs to regulate the target whose abnormal behavior causes a disease.

In both cases, traditional prediction methods build one classifier for each target (MHC molecule or drug target) separately. Using a kernel-based approach which casts the problem as predicting whether each pair, *e.g.* (peptide,MHC) or (molecule,target) interacts or not [1], we obtained significant prediction improvement in accuracy for the targets with few known binders. We have proposed some specific kernels for each problem, and shown that this approach improves the prediction accuracy for both the MHC [2] and drug discovery problems [3]. In [4], we propose some additional kernels for the GPCR case.

2 Clustered multi-task learning

Multi-task learning involves considering several related problems simultaneously, with the hope of improving performance by sharing information across these problems or "tasks".

A common strategy is to penalize the variance across the classification functions of all the tasks, which can help guide learning when little data is available.

In more realistic settings, it may be that certain inference problems are related but others are quite different. In such cases, penalizing the overall variance may harm the performance and one would like to penalize the variance only within clusters of related problems.

As these clusters are unknown a priori, we have proposed in [5] a criterion which penalizes the variance of functions within clusters, and optimize with respect to both the classification functions and clustering. Clustering being a non-convex problem, we have proposed a convex relaxation, which we show improved the prediction performances.

3 Structured priors for expression data analysis

A well known problem in bioinformatics is to predict the class of a tumour from gene expression measurements with microarrays, and simultaneously select a small number of genes to establish a predictive signature. Selecting a few genes that either belong to the same functional groups (where the groups are given a priori and may overlap *e.g.*, biological pathways) or tend to be connected to each other in a given biological network, may lead to increased interpretability of the signature and potentially to better performance when little data is available.

To this end, we proposed and studied in [6] a new penalty which generalizes the ℓ_1/ℓ_2 norm to overlapping groups, and cast the problem of selecting connected covariates in a graph as the problem of selecting a union of overlapping groups, with adequate definition of groups.

More generally, this method can be used in cases where either groups of covariates are given (potentially with overlap between the groups) and we wish to estimate a model whose support is a union of groups, or when a graph with covariates as vertices is given and we wish to estimate a model whose support contains covariates which tend to be connected to each other on the graph.

References

[1] J.-P. Vert and L. Jacob. Machine learning for in silico virtual screening and chemical genomics: New strategies. *Combinatorial Chemistry & High Throughput Screening*, 11(8):677–685, September 2008.

[2] L. Jacob and J.-P. Vert. Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics*, 24(3):358–366, Feb 2008.

[3] L. Jacob and J.-P. Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24(19):2149–2156, 2008.

[4] L. Jacob, B. Hoffmann, V. Stoven, and J.-P. Vert. Virtual screening of GPCRs: an *in silico* chemogenomics approach. *BMC Bioinformatics*, 9:363, 2008.

[5] L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems 21*, pages 745–752. MIT Press, 2009.

[6] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlaps and graph lasso. In *ICML'09 Proceedings of the 26th international conference on Machine learning*, 2009. To appear.