INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team ABC

Apprentissage et Biologie Computationnelle

Nancy - Grand Est

THEME BIO

Activity Report

2008

Table of contents

1. Team

Research Scientist

Yann Guermeur [Research Associate (CR) CNRS, Team Leader, HdR]

Faculty Member

Emmanuel Didiot [ATER, Nancy 2, until 08/2008]
Fabienne Thomarat [Associate Professor, INPL]

External Collaborator

Frédéric Bertrand [Associate Professor, ULP]
Alexander Bockmayr [Professor, Freie Universität Berlin, HdR]
Myriam Maumy [Associate Professor, ULP]

Administrative Assistant

Sylvie Thomas [Secretary (SAR) INRIA, from 02/2008 until 11/2008]

2. Overall Objectives

2.1. Overall Objectives

The aim of the ABC ("Apprentissage et Biologie Computationnelle", i.e., Machine Learning and Computational Biology) team is to develop the theory and practice of supervised and unsupervised learning. We focus on the theory of multi-class pattern recognition, deriving uniform convergence results which primarily deal with multi-class kernel machines such as multi-class SVMs (M-SVMs) [3]. Our applications are in the field of biological sequence processing. Precisely, our research themes can be summarized as follows:

- Derivation of bounds on the risk of classifiers
- Development of methods of model selection
- Specification and evaluation of multi-class support vector machines
- Protein secondary structure prediction
- Molecular phylogeny

A specificity of the team is its interdisciplinarity. Basically, our contributions belong to three fields: machine learning, bioinformatics and statistics. They are roughly uniformly distributed between these three fields.

2.2. Highlights

This year, we joined the "Pattern Analysis, Statistical Modelling and Computational Learning" (PASCAL 2) network of excellence.

3. Scientific Foundations

3.1. Introduction

The goal of classification is to assign objects to classes (also referred to as categories). There are two types of classification problems. Supervised learning qualifies configurations where the set of categories is known *a priori*. In ABC, we study this field in the context of statistical learning. Unsupervised learning, or classification in its strict sense, corresponds to situations when the set of categories is unknown.

3.2. Statistical learning

Statistical learning theory [17] is one of the fields of inferential statistics the bases of which have been established by V.N. Vapnik in the late sixties. The goal of this theory is to specify the conditions under which it is possible to "learn" from empirical data obtained by random sampling. Learning amounts to solving a problem of function or model selection. Basically, given a task characterized by a joint probability distribution P on a product space $\mathcal{X} \times \mathcal{Y}$, and a class of functions $\mathcal{F}$, of cardinality ordinarily infinite, the goal is to find a function $f \in \mathcal{F}$ with minimal *expected risk*, i.e., such that substituting $f(x)$ to y induces a minimal *loss* (the risk is the expectation of the loss function with respect to P). Training can thus be reformulated as an optimization problem. In many cases, the objective function used is related to the *capacity* of $\mathcal{F}$ [6]. The learning tasks considered belong to one of the three following domains: pattern recognition (discriminant analysis), function approximation (regression) and density estimation.

This theory considers more specifically two inductive principles. The first one, the empirical risk minimization (ERM) principle, consists in minimizing directly the training error. If the sample is small, one substitutes to this principle the structural risk minimization (SRM) inductive principle. It consists in minimizing an upper bound on the expected risk (generalization error), bound sometimes called a *guaranteed risk*, although it only holds true with high probability. This latter principle is implemented in the training algorithms of the support vector machines (SVMs), which currently constitute the state of the art for numerous problems of pattern recognition.

SVMs are *kernel machines* [15] conceived to compute indicator functions, to perform regression or to estimate densities. They have been introduced during the last decade by Vapnik and co-workers [11], [12], as nonlinear extensions of the maximal margin hyperplane [16]. Their main advantage is that they can avoid overfitting in the case where the size of the sample is small [17], [14].

3.3. Classification

Classification consists in clustering a set of objects in a finite number of categories unknown *a priori*. Objects are represented by their descriptions, which usually correspond to vectors of $\mathbb{R}^n$. Different types of methods can be used, gathered by literature into three main groups: mixture models, partitional clusterings and hierarchical clusterings. The objects can thus be clustered in various structures. While some methods only provide a partition of the objects, others generate hierarchical structures. We are primarily interested in hierarchical classifications, when the categories are arranged in a tree-structure.

Molecular phylogeny aims at identifying the evolutionary relationships between species, relationships which are represented by a tree. The descriptors of the species are usually derived from the sequence of a gene or a protein. To build a tree, different kinds of methods are used. Some of them are based on the characters (nucleotides or amino acids), the most common ones being the *maximum parsimony* method and the *maximum likelihood* method. The other methods are based on distances. Using a given criterion, the distance matrix methods aim at building a tree from a matrix of pairwise distances.

4. Application Domains

4.1. Molecular biology

Keywords: *Biology*, *biological sequence processing*, *structure prediction*.

Participants: Emmanuel Didiot, Yann Guermeur, Fabienne Thomarat.

Molecular biology is concerned with the study of three types of biological macromolecules: DNA, RNA, and proteins. Each of these molecules can initially be viewed as a string on a finite alphabet: DNA and RNA are nucleic acids made up of nucleotides A,C,G,T and A,C,G,U, respectively. Proteins are sequences of amino acids, which may be represented by an alphabet of 20 letters.

Molecular biology studies the information flow from DNA to RNA, and from RNA to proteins. In a first step, called *transcription*, a DNA string ("gene") is transcribed into messenger RNA (mRNA). In the second step, called *translation*, the mRNA is translated into a protein: each triplet of nucleotides encodes one amino acid according to the genetic code. The genes of eukaryotic cells are mostly composed of a succession of coding regions, called exons, and non-coding regions, called introns. During transcription, an intermediate step, the *splicing* process, is then necessary to remove the introns from the premessenger RNA. The remaining exons are concatenated yielding the mature RNA molecule. *Alternative splicing* is a regulatory mechanism by which variations in the incorporation of the exons into mRNA lead to the production of different forms of mature mRNAs and consequently to more than one related protein, or isoform.

Biological macromolecules are not just sequences of nucleotides or amino acids. Actually, they are complex three-dimensional objects. DNA shows the famous double-helix structure. RNA and proteins fold into complex three-dimensional structures, which depend on the underlying sequence. RNA is a single-stranded chain of nucleotides. However, a nucleotide in one part of the molecule can base-pair with a nucleotide in another part, following the Watson-Crick complementarity rules. This results in a folding of the molecule. The *secondary structure* of RNA indicates the set of base pairings in the three dimensional structure of the molecule. This information can be represented by a graph.

Proteins have several levels of structure. Above the primary structure (i.e., the sequence) is the *secondary structure*, which involves three basic types: α*-helices*, β*-sheets*, and aperiodic structure elements called *loops*. The spatial relationship of the secondary structures forms the tertiary structure. Several proteins can function together in a protein complex whose structure is referred to as the quaternary structure. A *domain* of a protein is a combination of secondary structure elements with some specific function. It contains an *active site* where an interaction with an external molecule may happen. A protein may have one or several domains.

The ultimate goal of molecular biology is to understand the *function* of biological macromolecules in the life of the cell. Function results from the *interaction* between different macromolecules, and depends on their structure. The overall challenge is to make the leap from sequence to function, through structure: the prediction of structure will help to predict the function.

Thanks to the huge number of gene and protein sequences available in the sequence databases, molecular phylogenetic analyses multiplied since a few decades. Molecular phylogeny [13] is the use of genes or protein sequences to gain information on the evolutionary history of organisms. By comparison of the sequence of a gene in different organisms, the evolutionary history of these sequences can be inferred. Based on the hypothesis that these sequences are orthologs (i.e., come from a same ancestral sequence by speciation events), the evolutionary history of the organisms can also be inferred and be represented by a tree.

5. Software

5.1. Multi-class Support Vector Machines

Participant: Yann Guermeur.

We have extended our M-SVM software, which was only implementing the M-SVM of Weston and Watkins so far, so that it implemented the two other main M-SVM models plus the M-SVM^2 (see for instance [3]). This new version of our application, used in intern for now, should be registred at the APP soon.

6. New Results

6.1. Prediction of subcellular targeting in heterokonts

Participant: Yann Guermeur.

The heterokonts are a particularly interesting group of eukaryotic organisms: they include many key species of planktonic and coastal algae and several important pathogens. To understand the biology of these organisms, it is necessary to be able to predict the subcellular localisation of their proteins but this is not straightforward, particularly in photosynthetic heterokonts which possess a complex chloroplast, acquired as the result of a secondary endosymbiosis. This is because the bipartite target peptides that deliver proteins to these chloroplasts can be easily confused with the signal peptides of secreted proteins, causing currently available algorithms to make erroneous predictions. In [8], we introduced HECTAR, a subcellular targeting prediction method which takes into account the specific properties of heterokont proteins. It is a hierarchical classifier whose architecture is a decision tree. At each node of the tree, the outputs of several classifiers are combined thanks to a SVM or a M-SVM.

6.2. Model selection for multi-class SVMs

Participant: Yann Guermeur.

Using a support vector machine requires to set two types of hyperparameters: the soft margin parameter C and the parameters of the kernel. To perform this model selection task, the method of choice is cross-validation. Its leave-one-out variant is known to produce an estimator of the generalization error which is almost unbiased. Its major drawback rests in its time requirement. To overcome this difficulty, several upper bounds on the leave-one-out error of the pattern recognition SVM have been derived. Among those bounds, the most popular one is probably the radius-margin bound. It applies to the hard margin pattern recognition SVM, and by extension to the 2-norm SVM. In [10], we derived a generalized radius-margin bound dedicated to the M-SVM of Lee, Lin and Wahba. We also introduced a quadratic loss M-SVM, the M-SVM^2, as a direct extension of the 2-norm SVM to the multi-class case.

6.3. Comparative study of multi-class support vector machines

Participant: Yann Guermeur.

Bi-class SVMs, introduced in bioinformatics at the end of the nineties, currently obtain state-of-the-art performance for numerous problems of biological sequence processing. Multi-class SVMs, of more recent conception, are progressively applied to these problems, especially in predictive structural biology. In [9], we described a comparative study of the performance of three multi-class SVMs in protein secondary structure prediction. The models involved were the one of Weston and Watkins, the one of Lee and co-authors and the M-SVM^2 (see Section 6.2).

7. Other Grants and Activities

7.1. Regional Initiatives

We participate in the "Génopole Strasbourg Alsace-Lorraine" together with the laboratory "Maturation des ARN et Enzymologie Moléculaire" (MAEM) and the "Institut de Génétique et de Biologie Moléculaire et Cellulaire" (IGBMC) in Strasbourg.

7.2. Actions Funded by the EC

We are members of the "Pattern Analysis, Statistical Modelling and Computational Learning" (PASCAL 2) network of excellence.

8. Dissemination

8.1. Serving the scientific community

Yann Guermeur has been a member of the program committee of the following conferences: "International Conference on Machine Learning" (ICML'08), "Apprentissage Artificiel et Fouille de Données" (AAFD'08) and "Maghrebian Conference on Information Technologies" (MCIT'08). He presented an invited communication at the conference AAFD'08 and the "Journées Modélisation Aléatoire et Statistique" (MAS) of the "Société de Mathématiques Appliquées et Industrielles" (SMAI). He was a member of the PhD jury of Christophe Magnan, Abderrahmane Boubezoul, Nicolas Garnier and Bernhard Gschloessl. For Bernhard Gschloessl, he also wrote a report. He is an expert for the ANR.

8.2. Teaching

Fabienne Thomarat is Associate Professor at the École Nationale Supérieure des Mines de Nancy / Institut National Polytechnique de Lorraine (engineering school, master of engineering school). She is in charge of one option (Bioinformatics) at the Department of Computer Science.

Yann Guermeur is in charge of the UE 3.105 entitled "Apprentissage statistique et fouille de données" of the M2P speciality "Génomique et Informatique" of the Master "Sciences de la Vie et de la Santé" (SVS), at the University Henri Poincaré. He gave two lectures at the "Summer School on Neural Networks in Classification, Regression and Data Mining" (NN 2008, http://www.isep.ipp.pt/nn/), one entitled "Multi-Class Support Vector Machines" and the other "Protein Secondary Structure Prediction with Multi-Class Support Vector Machines".

Emmanuel Didiot was ATER at the University Nancy 2 until August 2008.

9. Bibliography

Major publications by the team in recent years

[1] E. BALAS, A. BOCKMAYR, N. PISARUK, L. WOLSEY. *On unions and dominants of polytopes*, in "Mathematical Programming, Ser. A", vol. 299, 2004, p. 223-239.

[2] A. BOCKMAYR, J. N. HOOKER. *Constraint Programming*, in "12: Discrete Optimization", K. AARDAL, G. NEMHAUSER, R. WEISMANTEL (editors), Handbooks in Operations Research and Management Science, chap. 10, Elsevier, 2005, p. 559–600.

[3] Y. GUERMEUR. *SVM multiclasses, théorie et applications*, Habilitation à Diriger des Recherches, Université Henri Poincaré, 2007.

[4] Y. GUERMEUR. *VC Theory of Large Margin Multi-Category Classifiers*, in "Journal of Machine Learning Research", vol. 8, 2007, p. 2551–2594.

[5] Y. GUERMEUR, A. LIFCHITZ, R. VERT. *A kernel for protein secondary structure prediction*, in "Kernel Methods in Computational Biology", B. SCHÖLKOPF, K. TSUDA, J.-P. VERT (editors), The MIT Press, 2004, p. 193–206.

[6] Y. GUERMEUR, O. TEYTAUD. *Estimation et contrôle des performances en généralisation des réseaux de neurones*, in "Apprentissage Connexionniste", Y. BENNANI (editor), chap. 10, Hermès, 2006, p. 279–342.

[7] N. SAPAY, Y. GUERMEUR, G. DELÉAGE. *Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier*, in "BMC Bioinformatics", vol. 7, n$^{\mathrm{o}}$ 255, 2006.

Year Publications

Articles in International Peer-Reviewed Journal

[8] B. GSCHLOESSL, Y. GUERMEUR, J. COCK. *HECTAR: A method to predict subcellular targeting in heterokonts*, in "BMC Bioinformatics", vol. 9, n$^{\mathrm{o}}$ 393, 2008.

Articles in National Peer-Reviewed Journal

[9] Y. GUERMEUR. *Etude comparée des performances de SVM multi-classes en prédiction de la structure secondaire des protéines*, in "Revue des Nouvelles Technologies de l'Information", (to appear), 2008.

Research Reports

[10] E. MONFRINI, Y. GUERMEUR. *A Quadratic Loss Multi-Class SVM*, Technical report, LORIA, hal-00276700, 2008.

References in notes

[11] B. BOSER, I. GUYON, V. VAPNIK. *A training algorithm for optimal margin classifiers*, in "COLT'92", 1992, p. 144–152.

[12] C. CORTES, V. VAPNIK. *Support-Vector Networks*, in "Machine Learning", vol. 20, 1995, p. 273–297.

[13] J. FELSENSTEIN. *Inferring Phylogenies*, Sinauer, 2004.

[14] Y. GUERMEUR, H. PAUGAM-MOISY. *Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines*, in "Apprentissage Automatique", M. SEBBAN, G. VENTURINI (editors), (in French), Hermès, 1999, p. 109–138.

[15] B. SCHÖLKOPF, A. SMOLA. *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Cambridge, MA, 2002.

[16] V. VAPNIK. *Estimation of Dependences Based on Empirical Data.*, Springer-Verlag, N.Y., 1982.

[17] V. VAPNIK. *Statistical learning theory*, John Wiley & Sons, Inc., N.Y., 1998.