INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team ABC

Apprentissage et Biologie Computationnelle

Nancy - Grand Est

Theme :

Table of contents

1. Team

Research Scientist

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

Faculty Member

Martine Cadot [PRAG, UHP]
Fabien Lauer [Associate Professor, UHP, since 09/2009]
Fabienne Thomarat [Associate Professor, INPL]

External Collaborator

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

PhD Student

Rémi Bonidal [PhD Student, UHP, since 09/2009]
Hafida Bouziane-Chouarfia [Assistant Professor, USTO]

Administrative Assistant

Cécilia Claude [Secretary (SAR) INRIA]

Other

Gabriel Meurin [Master Student UHP, from 04/2009 until 09/2009]

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 support vector machines (M-SVMs) [2]. 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
- Robust data mining
- 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.

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 [24] 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 $\mathcal{F}$ [4]. 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 SVMs, which currently constitute the state of the art for numerous problems of pattern recognition. SVMs are *kernel machines* [22] 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 [16], [18], as nonlinear extensions of the maximal margin hyperplane [23]. Their main advantage is that they can avoid overfitting in the case where the size of the sample is small [24], [21].

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. This is especially the case of the methods used in molecular phylogeny, which generate 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.

Pairwise distance methods are also found in partitional clustering. Among them, spectral clustering has received much attention from the community. This framework is based on the embedding of the data into the eigenspace of a normalized form of the similarity matrix, which contains all the pairwise similarities. This embedding emphasizes the connectivity between the points and allows standard clustering algorithms to recover the correct labeling in difficult cases, where they would fail if applied directly to the data.

3.4. Robust data mining

Data mining consists of applying algorithms for producing models over the data. It can use algorithms from data analysis, statistics, pattern recognition, machine learning, or other fields like rules extraction (decision rules, association rules), as long as they deal with large data, which may reveal heterogeneous and non structured [19]: texts, biological and speech sequences, etc. To our eyes [17], Robust Data Mining consists in not doing any hypothesis about the data, such as normality, etc.; on the contrary, we use the principles of inferential statistics (randomization tests, ...) and machine learning (disjoint training and test sets) so as to guarantee the generalizing ability of the discovered models.

4. Application Domains

4.1. Molecular biology

Participants: Hafida Bouziane-Chouarfia, 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 [20] 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 programmed the M-SVM^2 introduced in [7]. This application, used in intern for now, should be registred at the APP soon.

6. New Results

6.1. Guaranteed risks

Participant: Yann Guermeur.

Bounds on the risk play a crucial role in statistical learning theory. They usually involve as capacity measure of the model studied the VC dimension or one of its extensions. In [6], introducing a class of generalized VC dimensions called γ-Ψ-dimensions, we computed the sample complexity of classifiers taking values in $\mathbb{R}^Q$. Then, we computed the Rademacher complexity of the M-SVMs. This provided us with a bound on their risk which is sharper than the one based on a γ-Ψ-dimension.

6.2. Model selection for multi-class SVMs

Participant: Yann Guermeur.

Using a SVM requires to set the values of 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 requirements in terms of computational time. 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 machine, and, by extension, to the 2-norm SVM. In [7], we have introduced a variant of the M-SVM of Lee, Lin and Wahba: the M-SVM^2. This quadratic loss machine can be seen as a direct extension of the 2-norm SVM to the multi-class case. For this machine, a generalized radius-margin bound was established in [11].

6.3. Spectral clustering of linear subspaces

Participant: Fabien Lauer.

Subspace separation is a particular case of clustering, where the data are distributed along multiple subspaces rather than around prototype centers. The increase in difficulty comes in part from the fact that multiple subspaces intersect, so that distributions of points belonging to different groups are more likely to overlap, often leading to indistinguishability. We showed in [12] that, for this problem, the dimension of the ambient space is crucial for separability, and that low dimensions chosen in prior work are not optimal. We suggested lower and upper bounds together with a data-driven procedure for choosing the optimal ambient dimension in a spectral clustering framework for linear subspace separation. Application of our approach to motion segmentation from tracked feature points uniformly outperformed a range of state-of-the-art methods both in terms of segmentation accuracy and computational speed.

6.4. Robust data mining and graphs

Participant: Martine Cadot.

In 2009, we focused our research interest on graph models for modeling relations between variables from data with two points of views : 1) in [9], we discussed how the statistical implications graph proposed by Régis Gras has been constructed to avoid the inconsistencies of a set of statistical implication rules for their uses in deduction. We compared this statistical model of the data to two alternative models: one is an algebraic model, the Galois lattice, the other one is probabilistic, i.e., Bayesian networks. 2) in [8] and [13], we proposed a statistical definition of neighborhood: our TourneBool randomization test processes an objects vs. variables binary table in order to establish which inter-variables relation is fortuitous, and which one is meaningful, out of any hypotheses on the underlying statistical distributions, but taking into account these empirical distributions. It ensues a robust and statistically validated graph of relations and "counter-relations". In [10], we used this technique for pruning a classification graph.

6.5. Protein secondary structure prediction

Participants: Hafida Bouziane-Chouarfia, Yann Guermeur.

Many neural networks have already been used for protein secondary structure prediction. It is known that this prediction can benefit from endowing the architecture of the networks with dedicated devices. In [14], we performed a comparative study of several multi-layer perceptrons dedicated to this task.

6.6. Text categorization in genomics

Participants: Martine Cadot, Gabriel Meurin.

We have worked with Michel Zitt (INRA-OST) and Alain Lelu (team KIWI) on evaluating the stability and generalization ability of a given partition in a bibliographic database. We used machine learning techniques for assessing these aspects in 1) an AKM 50 cluster partition of a 120 000 abstract genomics corpus and 2) a complex computer-based and expert validated partition between the genomic and non-genomic scientific domains, through one month of the Web of Science (68 000 abstracts) [15].

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

Martine Cadot is member of the editorial board of the International Journal of Data Mining, Modelling and Management (IJDMMM). She has been a member of the program committee of the following conferences: QDC'09, FDC'09 and eKNOW'10. She is a member of the SFdS.

Yann Guermeur has been a member of the program committee of the following conferences: the "XIIIth Applied Stochastic Models and Data Analysis International Conference" (ASMDA'09), the "Conférence Francophone sur l'Apprentissage Automatique" (CAp'09) and the "Journées Ouvertes en Biologie, Informatique et Mathématiques" (JOBIM'09). He was a member ("rapporteur") of the PhD jury of Karina Zapién Arreola. He is an expert for the ANR.

8.2. Teaching

Martine Cadot is PRAG in the department of Computer Science at the Université Henri Poincaré Nancy 1 where she teaches data mining.

Fabien Lauer is Associate Professor in the department of Computer Science at the Université Henri Poincaré Nancy 1.

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.

8.3. Theses, habilitations, academic duties

Yann Guermeur was "rapporteur" for the PhD thesis of Karina Zapién Arreola, at the INSA of Rouen.

9. Bibliography

Major publications by the team in recent years

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

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

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

[4] 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.

[5] 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^{o} 255, 2006.

Year Publications

Articles in International Peer-Reviewed Journal

[6] Y. GUERMEUR. *Sample Complexity of Classifiers Taking Values in $\mathbb{R}^Q$, Application to Multi-Class SVMs*, in "Communications in Statistics - Theory and Methods", vol. 39, n^{o} 3, 2010, p. 1–15.

[7] Y. GUERMEUR, E. MONFRINI. *A Quadratic Loss Multi-Class SVM for which a Radius-Margin Bound Applies*, in "Informatica", (submitted), 2009.

[8] A. LELU, M. CADOT. *Statistically valid links and anti-links between words and between documents: applying TourneBool randomization test to a Reuters collection.*, in "A paraitre (Springer)", 2009, p. 1-17, http://hal.archives-ouvertes.fr/hal-00429434/en/.

Articles in National Peer-Reviewed Journal

[9] M. CADOT. *Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayésiens et les treillis de Galois*, in "RNTI", vol. E, n^{o} 16, 2009, p. 223–250, http://hal.archives-ouvertes.fr/hal-00429437/en/.

International Peer-Reviewed Conference/Proceedings

[10] P. CUXAC, A. LELU, M. CADOT. *Suivi incrémental des évolutions dans une base d'information indexée : une boucle évaluation / correction pour le choix des algorithmes et des paramètres.*, in "2nd International Conference on Information Systems and Economic Intelligence (SIIE'2009)", 2009, http://hal.archives-ouvertes.fr/inria-00361652/en/.

[11] Y. GUERMEUR, E. MONFRINI. *Radius-Margin Bound on the Leave-One-Out Error of the LLW-M-SVM*, in "ASMDA'09", 2009, p. 517–521.

[12] F. LAUER, C. SCHNÖRR. *Spectral clustering of linear subspaces for motion segmentation*, in "IEEE Int. Conf. on Computer Vision (ICCV), Kyoto, Japan", 2009, p. 678–685.

[13] A. LELU, M. CADOT. *Graphes des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel*, in "Extraction et gestion de connaissance 2009 (EGC'09)", J.-G. GANASCIA (editor), Revue des Nouvelles Technologies de l'Information, vol. E-15, Cépaduès éditions, 2009, p. 367–378, http://hal.archives-ouvertes.fr/inria-00342751/en/.

Workshops without Proceedings

[14] H. BOUZIANE-CHOUARFIA, B. MESSABIH, Y. GUERMEUR, A. CHOUARFIA. *Approach Based on Artificial Neural Networks for Protein Secondary Structure Prediction*, in "NTICRI'09", 2009.

[15] M. CADOT, M. ZITT, G. MEURIN, A. LELU. *Investigating word interactions in texts. Application to text categorization in genomics*, in "First SaarLorLux Workshop on Systems Biology 2009, Computational, Structural and Medical Approaches for Systems Biology", 2009, http://hal.inria.fr/inria-00442395/en/.

References in notes

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

[17] M. CADOT. *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*, Ph. D. Thesis, Université de Franche-Comté, 2006.

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

[19] U. M. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH. *From data mining to knowledge discovery: an overview*, 1996, p. 1–34.

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

[21] 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.

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

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

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