Loria
Laboratoire lorrain de recherche
en informatique et ses applications

Lorraine Laboratory of
Research in Computer Science
and its Applications

# ACTIVITY REPORT 2010

TEAM

# ABC

## Apprentissage et Biologie Computationnelle

Research / Training / Transfer
in an international context

# Table of contents

# 1. Team

**Research scientist**

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

**Faculty members**

Martine Cadot [PRAG, Université Henri Poincaré (UHP)]
Fabien Lauer [Associate Professor, UHP]
Fabienne Thomarat [Associate Professor, INPL]

**Administrative staff**

Sylvie Musilli [Assistant, INPL]

**PhD students**

Hafida Bouziane-Chouarfia [Assistant Professor, USTO; defense planned in 2011]
Rémi Bonidal [UHP]

# 2. 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, implementation, and evaluation of multi-class support vector machines
- Robust data mining
- Statistical processing of biological sequences (protein secondary structure prediction, . . . )

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 theory

Statistical learning theory [33] is one of the fields of inferential statistics the bases of which have been laid 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 $\mathscr{X} \times \mathscr{Y}$, and a class of functions $\mathscr{F}$, of cardinality ordinarily infinite, the goal is to find a function $f \in \mathscr{F}$ 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. 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 can be related with Tikhonov's regularization theory [31]. In that framework, the objective function of the training algorithm incorporates a term related to the *capacity* of $\mathscr{F}$ [5]. The most famous example of implementation of this principle is the training algorithms of the SVMs. Those machines are *kernel machines* [30] conceived to compute indicator functions, to perform regression or to estimate densities. They have been introduced by Vapnik and his co-workers [25, 27], as nonlinear extensions of the maximal margin hyperplane [32]. Their main advantage is that they can avoid overfitting in the case where the size of the sample is small [33, 29]. Several M-SVMs have been proposed in literature (see [2] for a survey). There theory can be embeded in the theory of large margin multi-category classifiers [3, 4].

## 3.3. Robust data mining and classification

Data mining consists in applying algorithms for producing models over the data. It can use algorithms from the now deeply intersecting fields of data analysis, statistics, pattern recognition, machine learning, or other fields like rules extraction, as long as they deal with large data, which may reveal heterogeneous and non structured [28]: texts, biological and speech sequences, social and behavioural data. Robust Data Mining consists, to our eyes [26], 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 generalization capabilities of the discovered models.

Unsupervised classification consists in clustering a set of objects in a finite number of categories unknown *a priori*. 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. 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. A major open problem consists in determining the dimensionality of the reduced representation space in which the classification algorithms do operate.

Both data mining and unsupervised classification produce perspectives on the data structure that can be linked with one or more target variables to explain. After a data mining process, for example, one can extract a reduced set of decision rules out of a set of association rules. Regarding classification, the spectral embedding emphasizes the connectivity between data-points and allows standard learning algorithms to recover the correct labelling in difficult cases such as non-convex classes, where they would fail if applied directly to the data.

# 4. Application Domains

## 4.1. Molecular biology

**Participants**: Hafida Bouziane-Chouarfia, Yann Guermeur, Fabienne Thomarat.

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

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. Because in eukaryotic cells, most genes are composed of a succession of coding regions, called exons, and non-coding regions, called introns, this second step is generally preceded by an intermediate step, referred to as the splicing process, during which the introns are removed from the mRNA.

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. Proteins have several levels of structure. Above the primary structure (i.e., the sequence) is the *secondary structure*, which involves three basic types: $\alpha$*-helices*, $\beta$*-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.

# 5. Software

A major part of the research done in the ABC team, let it be theoretical or applied to biological sequence processing, gives rise to pieces of software.

## 5.1. M-SVM$^2$

**Participants**: Yann Guermeur.

On Septembre 13, 2010, the programs implementing the M-SVM$^2$ introduced in [9] have been registred at the APP under the registration number `IDDN.FR.001.370001.000.S.P.2010.000.30000`. They are governed by the CeCILL-B license.

## 5.2. MSVMpack

**Participants**: Yann Guermeur, Fabien Lauer [contact].

MSVMpack is an open source software package dedicated to the family of *multi-class* support vector machines. So far, four machines of this kind have been proposed in the literature. The current version of MSVMpack provides a unified implementation of all of them, while also offering a convenient basis to develop all the machines of the family. The package, available at the following address: `http://www.loria.fr/~lauer/MSVMpack/MSVMpack.html`, consists in a set of command-line tools with a callable library, both of which have been designed to favor the ease of use.

# 6. New Results

## 6.1. Multi-class support vector machines

### 6.1.1. Generic model of M-SVM

**Participants**: Yann Guermeur [contact], Fabien Lauer.

Roughly speaking, there is one main model of pattern recognition SVM, with several variants of lower popularity. On the contrary, among the different M-SVMs which can be found in literature, none is clearly favoured. On the one hand, they exhibit distinct statistical properties. On the other hand, multiple comparative studies between M-SVMs and decomposition methods have highlighted the fact that in practice, each model has its advantages and drawbacks. In [10], we have introduced a generic model of M-SVM. All the machines of this kind published so far appear as instances of this model. This definition makes it possible to devise new machines meeting specific requirements as well as to analyse globally the statistical properties of the M-SVMs.

### 6.1.2. Model selection for M-SVMs

**Participants**: Rémi Bonidal, Yann Guermeur.

Using an SVM requires to set the values of hyperparameters. 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 [9], we 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.

### 6.1.3. Ensemble methods for M-SVMs

**Participants**: Hafida Bouziane-Chouarfia, Yann Guermeur.

The multiplicity of the M-SVMs proposed in literature and the variety of their properties call for the evaluation of combinations of these models. In [19], we studied the combination of M-SVMs with linear ensemble methods (LEMs). The sample complexity of these combiners is low, which should prevent them from overfitting, and the outputs of two of them are estimates of the class posterior probabilities.

## 6.2. Protein secondary structure prediction

**Participants**: Hafida Bouziane-Chouarfia, Yann Guermeur, Fabienne Thomarat [contact].

Nowadays, there are quite a few secondary structure prediction methods with state-of-the-art prediction accuracy. Most of them share the cascade architecture introduced by Qian and Sejnowski: the outputs of a sequence-to-structure classifier are post-processed by a structure-to-structure classifier. In [23], we revisited this architecture by using all the M-SVMs published so far as base classifiers. Their outputs are post-processed by logistic regression models prior to being combined by the LEMs estimating the class posterior probabilities. The resulting prediction method achieves a state-of-the-art recognition rate with the advantage over the comparable methods that its sample complexity is far lower.

## 6.3. Robust data mining and classification

**Participants**: Martine Cadot.

We have uncovered the structure of data by adopting two different points of view:

1. getting rid of redundancy by selecting the sole relations that cannot be deduced from the others,

2. getting rid of "noise" by selecting only statistically significant relationships.

In line with the first point of view, an important part of our effort has focused on the extraction of data models: our MIDOVA method for decomposing/reconstructing a binary datatable by means of a set of necessary and sufficient characteristic itemsets gave rise to the publications [14] and [13, 21, 22, 17].

A second development line based on this method has given interesting results in the field of supervised classification: MIDOVA was used to select sets of itemsets related to the variable to explain, and we proved on various public test sets [7, 16, 15] that our results were as good as those of black box classifiers, while expliciting the combinations of variables grounding this performance. In [18] we interested in the stability of clustering results and the ability to describe each cluster by a minimum number of itemsets.

In line with the second point of view, our TourneBool randomization test allowed the extraction of statistically valid graphs out of a document-word matrix [12]: graph of links as well as anti-links between the words on the one hand, between the documents on the other hand. In [24], we showed that this method could also determine the intrisic dimension of a graph. This was empirically validated on two problems of spectral graph clustering.

In [8], we validated by TourneBool a graph which is involved in a potential incremental method for clustering a text stream.

## 6.4. Piecewise smooth function approximation

**Participants**: Fabien Lauer.

SVMs and other available kernel-based methods can readily solve nonlinear regression problems, with however the underlying assumption that the target function is smooth. On the other hand, learning a piecewise smooth function involving discontinuities at unknown locations is a much more difficult problem due to the intrinsic mixture of both classification and regression. In [11], we proposed a new method to solve this problem for the case of piecewise affine functions, and applied it in the context of hybrid system identification. This method is based on a continuous optimization framework involving only the real parameters of the model as variables, thus avoiding the use of discrete optimization. This allows the corresponding minimization problem to be solved efficiently even for very large data sets, which is a major advantage over other approaches from the literature.

In addition, the proposed approach was extended to piecewise smooth functions in [20] through the use of kernel functions. In this case, the critical issue is the number of parameters in the model, which typically grows with the number of data in kernel-based methods. In order to apply the method to large data sets, a support vector selection procedure, based on a maximum entropy criterion, was proposed to fix the model size prior to learning and limit the number of optimization variables.

# 7. Other Grants and Activities

## 7.1. Regional initiatives

We participate in the "Génopole Strasbourg Alsace-Lorraine" together with 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. Scientific animation

- 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: DBKDA'11, eKNOW'11, QDC'11, and AFDC'11. She is a member of the ACM, the SFC, and the SFdS.
- Yann Guermeur has been a member of the program committee of the following conferences: the "Stochastic Modeling Techniques and Data Analysis International Conference" (SMTDA'10), the "Conférence Francophone sur l'Apprentissage Automatique" (CAp'10) and the "Journées Ouvertes en Biologie, Informatique et Mathématiques" (JOBIM'10). He is an expert for the ANR.

## 8.2. Invited conferences

Y. Guermeur gave an invited talk at the workshop "Optimization and Learning: Theory, Algorithms and Applications".

## 8.3. Teaching

- Rémi Bonidal gave lectures at the Université Nancy 2.
- Martine Cadot is PRAG in the Department of Computer Science at the UHP where she teaches data mining to master (M2P) students. She is the advisor of many internships.
- Fabien Lauer is Associate Professor in the Department of Computer Science at the UHP where he teaches machine learning to master (M1) students.
- 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 the option bioinformatics at the Department of Computer Science.

## 8.4. Theses, habilitations, academic duties

- Yann Guermeur was a member ("rapporteur") of the HdR jury of Nicolas Wicker (Université de Strasbourg) and a member ("examinateur") of the PhD jury of Cécile Bonnard (Université Montpellier 2).

- Fabien Lauer was a member of the PhD jury of Bertrand Cornélusse (Université de Liège, Belgium).

# 9. Bibliography

## Major publications in recent years

[1] B. Gschloessl, Y. Guermeur, and J.M. Cock. HECTAR: A method to predict subcellular targeting in heterokonts. *BMC Bioinformatics*, 9(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. *Journal of Machine Learning Research*, 8:2551–2594, 2007.

[4] Y. Guermeur. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics - Theory and Methods*, 39(3):543–557, 2010.

[5] Y. Guermeur and O. Teytaud. Estimation et contrôle des performances en généralisation des réseaux de neurones. In Y. Bennani, editor, *Apprentissage Connexionniste*, chapter 10, pages 279–342. Hermès, 2006.

[6] N. Sapay, Y. Guermeur, and G. Deléage. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. *BMC Bioinformatics*, 7(255), 2006.

## Year publications

### Articles in International Peer-Reviewed Journal

[7] M. Cadot, A. Lelu, “Optimization of the representation space for qualitative data: A preliminary validation on classification problems”, *International Journal On Advances in Software*, 2010, (accepted).

[8] P. Cuxac, A. Lelu, M. Cadot, “Paving the way to next generation data-stream clustering: towards a unique and statistically valid cluster structure at any time step”, *International Journal of Data Mining, Modelling and Management (IJDMMM)*, 2010, (accepted).

[9] Y. Guermeur, E. Monfrini, “A Quadratic Loss Multi-Class SVM for which a Radius-Margin Bound Applies”, *Informatica*, 2010, (in press).

[10] Y. Guermeur, “A Generic Model of Multi-Class Support Vector Machine”, *International Journal of Intelligent Information and Database Systems (IJIIDS)*, 2010, (accepted).

[11] F. Lauer, G. Bloch, R. Vidal, “A continuous optimization framework for hybrid system identification”, *Automatica*, 2010, (accepted).

[12] A. Lelu, M. Cadot, “Statistically valid links and anti-links between words and between documents: applying TourneBool randomization test to a Reuters collection”, *Advances in Knowledge Discovery and Management (AKDM) 292*, 2010, p. 307–324, Collection : Studies in Computational Intelligence (Springer).

### Articles in National Peer-Reviewed Journal

[13] M. Cadot, D. El Hadj Ali, "Modélisation et extraction des liens complexes entre variables. Application à des données socio-économiques.", *Revue des Nouvelles Technologies de l'Information (RNTI)*, 2010, (accepted).

[14] A.-M. Masson, M. Cadot, V. Nahama-Fourguette, "Evolution de l'agressivité et de la psychopathologie de patients post-traumatiques au cours d'une thérapie cognitive", *Acta Psychiatrica Belgica 110*, 4, 2010, p. 21–28, (in press).

### International Peer-Reviewed Conference/Proceedings

[15] M. Cadot, A. Lelu, "A Novel Decomposition Algorithm for Binary Datatables: Encouraging Results on Discrimination Tasks", *in : RCIS 2010, Nice, France*, p. 57–68, 2010. (IEEE Computer Society).

[16] M. Cadot, A. Lelu, "Optimized Representation for Classifying Qualitative Data", *in : DBKDA 2010, Les Menuires, France*, p. 241–246, 2010. (IEEE Computer Society, Conference Publishing Service).

[17] M. Cadot, A. Lelu, "Representing interaction in multiway contingency tables: MIDOVA, CA and log-linear model", *in : Correspondence Analysis and Related Methods - CARME 2011*, Rennes France, 2010.

[18] M. Cadot, M. Zitt, G. Meurin, A. Lelu, "Robustesse des partitions de textes : une exploration autour de l'apport des motifs de mots", *in : JADT 2010, Rome, Italie*, 2010.

[19] Y. Guermeur, "Ensemble Methods of Appropriate Capacity for Multi-Class Support Vector Machines", *in : SMTDA'10*, 2010.

[20] F. Lauer, G. Bloch, R. Vidal, "Nonlinear hybrid system identification with kernel models", *in : 49th IEEE Int. Conf. on Decision and Control (CDC), Atlanta, GA, USA*, 2010.

### National Peer-Reviewed Conference/Proceedings

[21] M. Cadot, D. El Hadj Ali, "Liaisons complexes entre variables : les repérer, les valider", *in : AFDC'2010, Hammamet, Tunisie*, p. 115–126, 2010.

[22] D. El Hadj Ali, M. Cadot, "Estimation de l'impact de la décision du mariage sur la pauvreté des ménages tunisiens", *in : MASHS 2010, Lille, France*, p. 45–56, 2010.

[23] Y. Guermeur, F. Thomarat, "New cascade architecture for protein secondary structure prediction", *in : JOBIM'10*, 2010. (poster).

[24] A. Lelu, M. Cadot, "Espace intrinsèque d'un graphe et recherche de communautés", *in : MARAMI 2010, Toulouse, France*, p. 241–246, 2010.

## References in notes

[25] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.

[26] M. Cadot. *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association.* PhD thesis, Université de Franche-Comté, 2006.

[27] C. Cortes and V.N. Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

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

[29] Y. GUERMEUR AND H. PAUGAM-MOISY. Théorie de l'apprentissage de Vapnik et SVM, support vector machines. In M. Sebban and G. Venturini, editors, *Apprentissage Automatique*, pages 109–138. Hermès, 1999. (in French).

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

[31] A.N. TIKHONOV AND V.Y. ARSENIN. *Solutions of Ill-Posed Problems*. V.H. Winston & Sons, Washington, D.C., 1977.

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

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