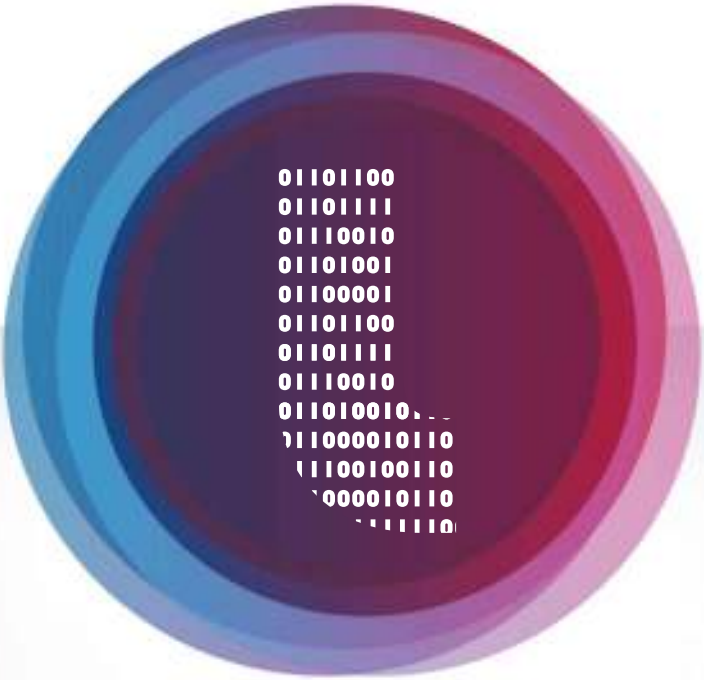
Loria
Laboratoire lorrain de recherche en informatique et ses applications

Lorraine Laboratory of
Research in Computer Science
and its Applications

ACTIVITY REPORT 2014



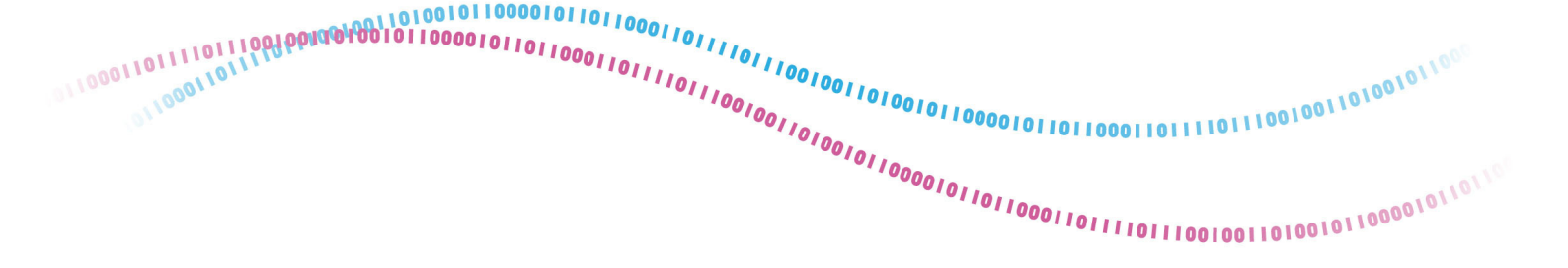
TEAM

ABC

Apprentissage et Biologie Computationnelle

Research / Training / Transfer
in an international context

Table of contents



1. Team

Research scientists

Yann Guermeur [Team Leader, Research Director, CNRS]

Faculty members

Fabien Lauer [Associate Professor, UL]

Administrative staff

Christelle Lévêque [Assistant, CNRS]

Technical staff

Emmanuel Didiot [Research Engineer, CNRS]
Khadija Musayeva [Expert Engineer, UL, from October until December 2014]

PhD student

Hafida Bouziane-Chouarfia [Assistant Professor, USTO; defended in November 2014]
Mounia Hendel [Assistant Professor, USTO]

Internships

Khadija Musayeva [UL, from March until September 2014]

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
- Specification, implementation, and evaluation of kernel methods for regression
- Statistical processing of biological sequences (biological sequence segmentation, ...)

A specificity of the team is its interdisciplinarity. Basically, our contributions belong to three fields: machine learning, bioinformatics and statistics.

2.1. Highlights of the year

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 [23] 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 [21]. In that framework, the objective function of the training algorithm incorporates a term related to the *capacity* of $\mathscr{F}$ [18]. The most famous example of implementation of this principle is provided by the training algorithms of the SVMs. Those machines are *kernel machines* [19] conceived to compute indicator functions, to perform regression or to estimate densities. They have been introduced by Vapnik and his co-workers [14, 15], as nonlinear extensions of the maximal margin hyperplane [22]. Their main advantage is that they can avoid overfitting in the case where the size of the sample is small [23]. Several M-SVMs have been proposed in literature (see [2] for a survey). Their theory can be embeded in the theory of large margin multi-category classifiers [16, 17].

4. Application Domains

4.1. Molecular biology

Participants: Hafida Bouziane-Chouarfia, Emmanuel Didiot, Yann Guermeur, Fabien Lauer.

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: 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.

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. MSVMpack

Participants: Emmanuel Didiot, Fabien Lauer [contact].

This package implementing all the M-SVMs published so far [5, 2] is distributed both via our website (`http://www.loria.fr/~lauer/MSVMpack/`) and via the Machine Learning Open Source Software website (`http://mloss.org`), where the number of downloads is now close to 4300. Updates are regularly released.

Since July 2014, a Windows port of MSVMpack is available that increases its popularity, especially in the community of biologists.

6. New Results

6.1. Multi-class pattern recognition

Participants: Yann Guermeur.

The authors of [20] present M-SVMs which they have introduced in recent years: multiobjective M-SVMs. Those machines are based on the same functional class as that of the standard M-SVMs [2]. They differ in the nature of the learning problem, which is no longer a standard optimization problem (convex quadratic programming problem), but a multiobjective optimization problem (taking the form of a second-order cone programming problem). The aim is to maximize exactly all geometric margins, so as to improve generalization performance. This performance is assessed empirically, through experiments performed on data sets from the UCI benchmark repository. In our comments [7], we make use of the latest results of the statistical theory of large margin multi-category classifiers to study the connection between the (width of the) geometric margins and the generalization performance.

6.2. Regression and sparse optimization

Participants: Fabien Lauer.

6.2.1. Piecewise smooth regression

We extended in [11] the recent approach of Ohlsson and Ljung for piecewise affine regression to the nonlinear case while taking a clustering point of view. In this approach, the problem is cast as the minimization of a

convex cost function implementing a trade-off between the fit to the data and a sparsity prior on the number of pieces. In particular, we considered the nonlinear case of piecewise smooth regression without prior knowledge on the type of nonlinearities involved. This was tackled by simultaneously learning a collection of local models from a reproducing kernel Hilbert space via the minimization of a convex functional, for which we proved a representer theorem that provides the explicit form of the solution. An example of application to piecewise smooth system identification showed that both the mode and the nonlinear local models can be accurately estimated.

6.2.2. DC programming for switching linear regression

In [10], we focused on switching linear regression in the large-scale setting with both numerous data and many parameters to learn. We considered the minimum-of-error framework with a quadratic loss function, in which an objective function based on a sum of minimum errors with respect to multiple submodels is to be minimized. We proposed a new approach to the optimization of this nonsmooth and nonconvex objective function, which relies on Difference of Convex (DC) functions programming. In particular, we formulated a proper DC decomposition of the objective function, which allowed us to derive a computationally efficient DC algorithm. Numerical experiments showed that the method can efficiently and accurately learn switching models in large dimensions and from many data points.

6.2.3. Iteratively reweighting scheme for sparse recovery and switching regression

Motivated by a sparse optimization approach to switching regression, we investigated in [9] the generic problem of recovering sparse solutions of underdetermined systems of linear equations, also often referred to as compressive sensing in the signal processing/information theory literature. More precisely, we focused on the associated convex relaxation where the ℓ_1-norm of the vector of variables is minimized and proposed a new iteratively reweighted scheme in order to improve the conditions under which this relaxation provides the sparsest solution. We proved the convergence of the new scheme and derived sufficient conditions for the convergence towards the sparsest solution. Experiments showed that the new scheme significantly improves upon the previous approaches for the generic problem on the one hand and for switching regression on the other hand.

6.2.4. Nonlinear sparse optimization

The sparse optimization community recently showed interest for extensions of results from linear sparse recovery as described above to the nonlinear case. In [8], we considered the problem of finding sparse solutions to systems of polynomial equations possibly perturbed by noise. In particular, we showed how these solutions can be recovered from group-sparse solutions of a derived system of linear equations. Then, two approaches were considered to find these group-sparse solutions. The first one was based on a convex relaxation resulting in a second-order cone programming formulation which can benefit from efficient reweighting techniques for sparsity enhancement. For this approach, sufficient conditions for the exact recovery of the sparsest solution to the polynomial system were derived in the noiseless setting, while stable recovery results were obtained for the noisy case. Though lacking a similar analysis, the second approach provided a more computationally efficient algorithm based on a greedy strategy adding the groups one-by-one. With respect to previous work, the proposed methods recover the sparsest solution in a very short computing time while remaining at least as accurate in terms of the probability of success. This probability was empirically analyzed to emphasize the relationship between the ability of the methods to solve the polynomial system and the sparsity of the solution.

6.3. Biological sequence segmentation

Participants: Hafida Bouziane-Chouarfia, Emmanuel Didiot, Yann Guermeur, Fabien Lauer [contact]

Khadija Musayeva.

We are interested in problems of bioinformatics which can be stated as follows: given a biological sequence and a finite set of categories, split the sequence into consecutive segments each assigned to a category different from those of the previous and next segments. Many problems of central importance in biology fit in this framework, such as protein secondary structure prediction, solvent accessibility prediction, splice site / alternative splicing prediction or the search for the genes of non-coding RNAs, to name just a few. Our aim is to devise a global solution for the whole class of these problems. On the one hand, it should be generic enough to allow a fast implementation on any instance. On the other hand, it should be flexible enough to make it possible to incorporate efficiently the knowledge available for a specific problem, so as to obtain state-of-the-art performance.

The solution advocated by the ABC team is based on a hybrid architecture that combines discriminative and generative models in the framework of a modular and hierarchical approach. In [12], we introduce the latest version of the application. It is evaluated on one specific task: protein secondary structure prediction. For this task, the use of input data derived from two types of position-specific scoring matrices (PSSMs) is investigated in [13]. The fusion of knowledge sources induces a gain in prediction accuracy which is statistically significant.

7. Collaborations and Contracts

7.1. Contracts with Industry

7.2. National Initiatives

7.3. International Initiatives

8. Dissemination

8.1. Scientific Animation

Yann Guermeur has been a member of the program committee of the following conferences: "European Conference on Artificial Intelligence" (ECAI) 2014, "IAPR International Conference on Pattern Recognition in Bioinformatics" (PRIB) 2014, and the "Conférence Francophone sur l'Apprentissage Automatique" (CAp) 2014.

8.2. Committees memberships

Yann Guermeur was a member of the PhD jury of Nguyen Manh Cuong (Université de Lorraine) and the HDR jury of Khalid Benabdeslem (Université Lyon 1). He was also referee for the PhDs of Rohit Babbar (Université de Grenoble) and Nicolas Jung (Université de Strasbourg).

8.3. Vulgarization

8.4. Invited Conferences

8.5. Teaching

Fabien Lauer is Associate Professor in the Department of Computer Science at the UL where he teaches machine learning to master (M1) students.

9. Bibliography

Major publications in recent years

[1] R. BONIDAL, S. TINDEL, AND Y. GUERMEUR. Model selection for the ℓ_2-SVM by following the regularization path. *Transactions on Computational Collective Intelligence*, Vol. XIII (LNCS 8342):83–112, 2014.

[2] Y. GUERMEUR. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 6(6):555–577, 2012.

[3] Y. GUERMEUR. Combining multi-class SVMs with linear ensemble methods that estimate the class posterior probabilities. *Communications in Statistics - Theory and Methods*, 42(16):2311–2330, 2013.

[4] F. LAUER, G. BLOCH, AND R. VIDAL. A continuous optimization framework for hybrid system identification. *Automatica*, 47(3):608–613, 2011.

[5] F. LAUER AND Y. GUERMEUR. MSVMpack: a multi-class support vector machine package. *Journal of Machine Learning Research*, 12:2293–2296, 2011.

[6] V.L. LE, G. BLOCH, AND F. LAUER. Reduced-size kernel models for nonlinear hybrid system identification. *IEEE Transactions on Neural Networks*, 22(12):2398–2405, 2011.

Year publications

Articles in International Peer-Reviewed Journal

[7] Y. GUERMEUR, "Comments on: Support vector machines maximizing geometric margins for multi-class classification", *TOP 22*, 3, 2014, p. 844–851.

[8] F. LAUER, H. OHLSSON, "Finding sparse solutions of polynomial systems of equations via group sparsity optimization", *Journal of Global Optimization*, 2014, (to appear).

[9] V. L. LE, F. LAUER, G. BLOCH, "Selective ℓ_1 minimization for sparse recovery", *IEEE Transactions on Automatic Control 59*, 11, 2014, p. 3008–3013.

[10] T. PHAM DINH, H. A. LE THI, H. M. LE, F. LAUER, "A difference of convex functions algorithm for switched linear regression", *IEEE Transactions on Automatic Control 59*, 8, 2014, p. 2277–2282.

International Peer-Reviewed Conference/Proceedings

[11] F. LAUER, G. BLOCH, "Piecewise smooth system identification in reproducing kernel Hilbert space", *in : Proc. of the 53rd IEEE Conf. on Decision and Control (CDC), Los Angeles, CA, USA*, p. 6498–6503, 2014.

Scientific Books (or Scientific Book chapters)

[12] Y. GUERMEUR, F. LAUER, "A Generic Approach to Biological Sequence Segmentation Problems, Application to Protein Secondary Structure Prediction", *in : Pattern Recognition in Computational Molecular Biology: Techniques and Approaches*, M. Elloumi, C. Iliopoulos, J. Wang, and A. Zomaya (editors), Wiley, 2014, (in press).

Other Publications

[13] K. MUSAYEVA, *Statistical Learning for Biological Sequence Segmentation*, Mémoire, Master Informatique de l'UL, 2014.

References in notes

[14] B. BOSER, I. GUYON, AND V. VAPNIK. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.

[15] C. CORTES AND V.N. VAPNIK. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

[16] Y. GUERMEUR. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8:2551–2594, 2007.

[17] 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.

[18] 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.

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

[20] K. TATSUMI AND T. TANINO. Support vector machines maximizing geometric margins for multi-class classification. *TOP*, 22(3):815–840, 2014.

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

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

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