Loria
Laboratoire lorrain de recherche
en informatique et ses applications

Lorraine Laboratory of
Research in Computer Science
and its Applications

ACTIVITY REPORT 2012

TEAM

ABC

Apprentissage et Biologie Computationnelle

Research / Training / Transfer
in an international context

Table of contents

1. Team

Research scientist

Yann Guermeur [Team Leader, Research Director, CNRS]

Faculty members

Martine Cadot [PRAG, Université de Lorraine (UL)]
Fabien Lauer [Associate Professor, UL]
Hoai An Le Thi [Professor, UL, since 9/2012]
Fabienne Thomarat [Associate Professor, UL]

Administrative staff

Sylvie Musilli [Assistant, UL]

PhD students

Rémi Bonidal [UL; defense planned in 2013]
Hafida Bouziane-Chouarfia [Assistant Professor, USTO; defense planned in 2013]
Mounia Hendel [Assistant Professor, USTO]
Edouard Klein [UL; defense planned in 2013]

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 [39] 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 [37]. 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 provided by the training algorithms of the SVMs. Those machines are *kernel machines* [36] conceived to compute indicator functions, to perform regression or to estimate densities. They have been introduced by Vapnik and his co-workers [27, 29], as nonlinear extensions of the maximal margin hyperplane [38]. Their main advantage is that they can avoid overfitting in the case where the size of the sample is small [39, 32]. 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, 12].

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 [30]: texts, biological and speech sequences, social and behavioural data. Robust Data Mining consists, to our eyes [28], 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 the 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.

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.

3.4. Optimization in Machine Learning and Data Mining (MLDM)

Models and optimization methods are proving to be vital in designing algorithms to extract essential knowledge from huge volumes of data: most methods in MLDM use optimization, or are themselves optimization algorithms. The interaction between optimization and MLDM is one of the most important developments in modern computer science. The major difficulty of the development of optimization for MLDM lies in the non convexity of the associate optimization model on one hand, and the very large dimension of this model on the other hand.

To overcome this difficulty we need sophisticated techniques and high-performance algorithms based on solid theoretical foundations and statistics. Based on the powerful arsenal of convex analysis, DC (Difference of Convex functions) programming and DCA (DC Algorithm) ([35, 34, 33] and reference therein) are among the few nonconvex optimization approaches that can meet these requirements. MLDM represents a mine of optimization problems that are almost all DC programs for which appropriate resolutions should use DC programming and DCA. During the last decade DC programming and DCA have been successfully applied to modeling and resolution of many problems in MLDM (including variable selection in classification, improved techniques for boosting, ϕ-learning, clustering, ... see the incomplete list of references in [33]). It is important to emphasize the impact of DC programming and DCA in combinatorial optimization that becomes increasingly important in MLDM. Being an effective, robust, and scalable approach, DCA should continue to play a vital role in meeting the challenges of MLDM.

4. Application Domains

4.1. Molecular biology

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

4.2. Image analysis

Participants: Hoai An Le Thi.

Compressed Sensing or Compressive Sensing (CS) is an emerging area having signifcant interest in data analysis. It can be used for compressing higher dimensional data sets to lower dimensional ones for data analysis, signal processing and feature selection applications. Since CS was introduced, it has been applied in various fields including radar imaging, signal extraction, aerial laser scanning, medical imaging, surface metrology, through wall radar imaging, space based imaging, ground penetrating radar imaging in archeology, geophysics, oil-exploration, landmine detection, forensics, and civil engineering.

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. MSVMpred2

Participants: Rémi Bonidal, Yann Guermeur, Fabien Lauer, Fabienne Thomarat [contact].

To tackle segmentation problems on biological sequences, we advocate the use of a hybrid architecture combining discriminant and generative models in the framework of a hierarchical approach. The discriminant models provide estimates of the class posterior probabilities which ared used to derive the emission probabilities of the generative model. We named MSVMpred a three-layer cascade of classifiers implementing the bottom part of the hybrid architecture. Multi-class support vector machines and neural networks provide a set of initial predictions. These predictions are post-processed by classifiers estimating the class posterior probabilities. At last, the outputs of these second-level classifiers are combined by means of a convex combination. The instance of MSVMpred dedicated to protein secondary, named MSVMpred2, is available online at the following address: `http://plateforme-mbi.loria.fr/MSVMpred/`.

5.2. MSVMpack

Participants: Fabien Lauer.

Updates of this package implementing all the M-SVMs published so far [6] are regularly released at the following address: `http://www.loria.fr/~lauer/MSVMpack/`.

5.3. *k*-LinReg

Participants: Fabien Lauer.

k-LinReg is an open source software dedicated to switched linear regression with large data sets. It is based on a simple *k*-means like algorithm described in [14] which efficiently provides a satisfactory solution to a difficult and nonconvex problem. Applications of this software include for instance switched linear and piecewise affine dynamical system identification. *k*-LinReg is available both as a platform-independent Matlab implementation and a parallel implementation in C for Linux. The software is available online at `http://www.loria.fr/~lauer/klinreg/`.

6. New Results

6.1. Support vector machines

6.1.1. Generic model of M-SVM

Participants: Yann Guermeur.

Roughly speaking, there is one main model of pattern recognition support vector machine, with several variants of lower popularity. On the contrary, among the different multi-class support vector machines 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 multi-class support vector machines and decomposition methods have highlighted the fact that in practice, each model has its advantages and drawbacks. In [12], we have introduced a generic model of multi-class support vector machine. It provides the first unifying definition of all the machines of this kind published so far. This contribution makes it possible to devise new machines meeting specific requirements as well as to analyse globally the statistical properties of the multi-class support vector machines.

6.1.2. Model selection for the ℓ_2-SVM

Participants: Rémi Bonidal, Yann Guermeur.

For a support vector machine, model selection consists in selecting the kernel function, the values of its parameters, and the amount of regularization. To set the value of the regularization parameter, one can minimize an appropriate objective function over the regularization path. A priori, this requires the availability of two elements: the objective function and an algorithm computing the regularization path at a reduced cost. The literature provides us with several upper bounds and estimates for the leave-one-out cross-validation error of the ℓ_2-SVM. However, no algorithm was available so far for fitting the entire regularization path of this machine. In [9], we introduced the first algorithm of this kind. It is involved in the specification of new methods to tune the corresponding penalization coefficient, whose objective function is a leave-one-out error bound or estimate. From a computational point of view, these methods appear especially appropriate when the Gram matrix is of low rank. A comparative study involving state-of-the-art alternatives provides us with an empirical confirmation of this advantage.

6.1.3. Combination of M-SVMs and estimation of the class posterior probabilities

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

In [16], we evaluated MSVMpred both as a stand alone classifier, i.e., according to the recognition rate, and as part of a hybrid architecture, i.e., with respect to the quality of the probability estimates.

Roughly speaking, there is one main model of pattern recognition support vector machine, with several variants of lower popularity. On the contrary, among the different multi-class support vector machines which can be found in the literature, none is clearly favoured. On the one hand, they exhibit distinct statistical properties. On the other hand, multiple comparative studies between multi-class support vector machines and decomposition methods have highlighted the fact that each model has its advantages and drawbacks. These observations call for the evaluation of combinations of multi-class support vector machines. In [13], we studied the combination of multi-class support vector machines with linear ensemble methods. Their sample complexity is low, which should prevent them from overfitting, and the outputs of two of them are estimates of the class posterior probabilities.

6.1.4. Applications of M-SVMs

Participants: Yann Guermeur, Hoai An Le Thi.

In [8], we addressed the problem of multi-class classification of skin pre-cancerous stages based on bimodal spectroscopic features combining spatially resolved AutoFluorescence (AF) and Diffuse Reflectance (DR) measurements. A new hybrid method to extract and select features is presented. It is based on Discrete Cosine Transform (DCT) applied to AF spectra and on Mutual Information (MI) applied to DR spectra. The classification is performed by means of a multi-class SVM: the M-SVM^2. Its performance is compared with the one of the One-Versus-All (OVA) decomposition method involving bi-class SVMs as base classifiers. The results of this study show that bimodality and the choice of an adequate spatial resolution allow for a significant increase in diagnostic accuracy. This accuracy can get as high as 81.7% when combining different distances in the case of bimodality.

In [20] we addressed the Intrusion Detection Systems (IDSs) problem which has an important role in network security. The main tasks for an IDS are to observe behaviors, identify the intrusions, and give the necessary warnings. These tasks are performed through the examination of network traffic to monitor signs of different malicious activities. M-SVMs are widely used for network intrusion detection. In that context, the main drawback of the decomposition methods (involving binary classifiers) rests in the fact that they cannot capture correlations between the different categories. In [20], we compared the M-SVM of Weston and Watkins with a mixed norm variant introduced in [31]. Both models were tested on the KDD Cup 1999 dataset which is very popular in the researches on network intrusion detection. The numerical results highlight their efficiency for the problem of interest.

6.2. Protein secondary structure prediction

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

Most of the state-of-the-art methods for protein seconday structure prediction are complex combinations of discriminant models. They apply a local approach of the prediction which is known to induce a limit on the expected prediction accuracy. A priori, the use of generative models should make it possible to overcome this limitation. However, among the numerous hidden Markov models which have been dedicated to this task over more than two decades, none has come close to providing comparable performance. A major reason for this phenomenon is provided by the nature of the relevant information. Indeed, it is well known that irrespective of the model implemented, the prediction should benefit significantly from the availability of evolutionary information. Currently, this knowledge is embedded in position-specific scoring matrices which cannot be processed easily with hidden Markov models. With this observation at hand, the next significant advance should come from making the best of the two approaches, i.e., using a generative model on top of discriminant models. In [25] we introduced the first hybrid architecture of this kind with state-of-the-art performance. The conjunction of the two levels of treatment makes it possible to optimize the recognition rate both at the residue level and at the segment level.

6.3. Robust data mining and classification

6.3.1. Development of a classification method

Participants: Martine Cadot.

Basically, MIDOVA lists the relevant combinations of K boolean variables, thus giving rise to an expansion of the original set of variables, well-fitted to for a number of data mining tasks. MIDOVA takes into account the presence as well as the absence of items. The building of level-k itemsets starting from level-k-1 ones relies on the concept of residue, which entails the potential of an itemset to create higher-order non-trivial associations. In [10] we have assessed the value of such a representation by presenting an application to three well-known classification tasks. The resulting success proves that our objective of extracting the relevant interactions hidden in the data, and only these ones, has been hit.

6.3.2. Development of a method for dimensionality reduction

Participants: Martine Cadot.

Laplacian low-rank approximations are much appreciated in the context of graph spectral methods and Correspondence Analysis. In [23] we addressed the problem of determining the dimensionality K* of the relevant eigenspace of a general binary datatable by a statistically well-founded method. In this paper, we have proposed 1) a general framework for graph adjacency matrices and any rectangular binary matrix, 2) a randomization test for fixing K*. Experimental evidence was provided by both artificial and real-world data.

6.3.3. Robust data mining for complex data

Participants: Martine Cadot.

To extract knowledge for complex data, usual data mining has to be combined with inferential statistics. We have studied 3 types of complex data: attitudes, text, and speech. [24] deals with attitudes. We have measured the effects of communication of uncertain knowledge. Uncertainty communication is assumed by some to increase public trust in science and policy-makers, by others to produce public panic. We have used focus groups for getting insights about this assumption and more generally about peoples' attitudes following uncertainty communication regarding the controversy on the effects of endocrine disrupters (EDs) on human male fertility. In [26], we have addressed the problem of thematic change detection in text by means of a collaboration of unsupervised and supervised methods. At last, speech data have been processed in [17, 18], precisely sequences of radiographs of a person talking. For several reasons it is difficult to analyze these data. We focused on the complexity of movements of the articulators during speech (tongue, jaw, etc.) We presented the extraction of articulatory models of speech from the data without adding a priori knowledge, using an ensemble of methods of data mining. These models have revealed the organization of articulatory structures in both the spatial dimension and the temporal dimension. Their comparison with expected movements of the articulators by the expert has been successful and invites us to continue along this path.

6.4. Regression

6.4.1. Learning piecewise smooth functions

Participants: Fabien Lauer, Hoai An Le Thi.

A number of efficient machine learning tools, including SVMs, exist to learn smooth functions with high accuracy from a finite data sample. However, when the target function is nonsmooth, the accuracy of these approaches becomes less satisfactory. In [19], we proposed a learning framework and a set of algorithms for nonsmooth regression, i.e., for learning piecewise smooth target functions with discontinuities in the function itself or the derivatives at unknown locations. In the proposed approach, the model belongs to a class of smooth functions, more precisely, it is built as a kernel expansion over the training data. Though constrained to be globally smooth, the trained model can have very large derivatives at particular locations to approximate the nonsmoothness of the target function. This is obtained through the definition of new regularization terms which penalize the derivatives in a location-dependent manner and training algorithms in the form of convex optimization problems. This constitutes a major advance over previous approaches which mostly rely on nonconvex optimization techniques, which are bound to yield suboptimal solutions, but also paves the way for considering piecewise smooth regression in the classical framework of penalized empirical risk minimization.

6.4.2. Probability of success for switched linear regression

Participants: Fabien Lauer.

In switched linear regression, the data is assumed to be generated by a collection of linear models. In this problem, the major difficulty is due to the fact that the data comes unlabeled, i.e., the index of the linear model that generated a particular data point is unknown. In [14], we analyzed *k*-LinReg, a straightforward and easy to implement algorithm in the spirit of *k*-means for the nonconvex optimization problem at the core of switched linear regression, and focused on the question of its accuracy on large data sets and its ability to reach global optimality. To this end, we emphasized the relationship between the sample size and the probability of obtaining a local minimum close to the global one with a random initialization. This was achieved through the estimation of a model of the behavior of this probability with respect to the problem dimensions. This model can then be used to tune the number of restarts required to obtain a global solution with high probability. Experiments showed that the model can accurately predict the probability of success and that, despite its simplicity, the resulting algorithm can outperform more complicated approaches in both speed and accuracy.

6.5. Nonconvex optimization techniques for MLDM

6.5.1. Learning with sparsity by DC programming and DCA

Participants: Hoai An Le Thi.

Sparse modeling of data has been developed over the years and is becoming increasingly essential in areas involving very large amounts of data. It eliminates the redundancy of data and considers only the most relevant features by removing unnecessary ones. Its applications are numerous: variable selection in classification, compressed sensing, portfolio management in finance, ... This involves NP-hard optimization problems dealing with the zero norm (the number of nonzero coefficients) of vectors representing data (in the objective or constraints). This theme draws the attention of many researchers in recent years, largely because of many topics in MLDM is related to the sparsity of data. Conventional methods in optimization are generally unapplicable or ineffective for this problem. Faced with this situation, approximation methods have been proposed to replace the zero norm in the original optimization problem by its (convex or nonconvex) approximation to obtain more suitable resolution methods. It is interesting to note that all approximations of the zero norm proposed in the literature are DC functions and the best existing algorithms in the group "non-convex approximation" are based on DCA. The development of new models and algorithms to optimization problems dealing with the zero norm remains a challenge for researchers in optimization and MLDM communities.

Sparsity of a classifier is a desirable condition for high dimensional data and large sample sizes. In [11] we investigate the two complementary notions of sparsity for binary classification: sparsity in the number of features and sparsity in the number of examples. Several different losses and regularizers are considered: the hinge loss and ramp loss, and ℓ_2, ℓ_1, approximate ℓ_0, and capped ℓ_1 regularization. We propose two new objective functions that further promote sparsity, corresponding to the ramp loss versions of approximate ℓ_0 and capped ℓ_1 regularization. We derive difference of convex functions algorithms (DCA) for solving these novel non-convex objective functions. We also propose an efficient DCA for optimizing the recently studied capped ℓ_1 regularizer under hinge loss. The proposed algorithms are shown to converge in a finite number of iterations to a local minimum. Using simulated data and several datasets from the UCI machine learning repository, we empirically investigate the fraction of features and examples required by the different classifiers.

As an application of sparsity, we consider in [22] the problem of signal recovery which is formulated as a ℓ_0-minimization problem. Using two appropriate continuous approximations of ℓ_0-norm, we reformulate the problem as a DC (Difference of Convex functions) program. DCA (DC Algorithm) is then developed to solve the resulting problems. Computational experiments on several datasets show the efficiency of our methods.

6.5.2. Classification of massive datasets by DC programming and DCA

Participants: Hoai An Le Thi.

1) Block Clustering based on DC programming and DCA

In [15] we investigate the DC (Difference of Convex functions) programming and DCA (DC Algorithm) to solve the Block clustering problem in the continuous framework, which traditionally requires solving a hard combinatorial optimization problem. DC reformulation techniques and exact penalty in DC programming are developed to build an appropriate equivalent DC program of the Block clustering problem. They lead to an elegant explicit DCA scheme for the resulting DC program. Computational experiments show the robustness and efficiency of the proposed algorithm and its superiority over standard algorithms in the literature such as Two-mode K-means, Two-mode Fuzzy Clustering, Block Classification EM.

2) DC formulations and DCA for clustering with weights on the variables

The performance of clustering algorithms can be significantly degraded if many attributes (variables) are irrelevant. Hence the interest of clustering models with weights on variables, especially for massive data. These models are more complicated than standard clustering formulations. In [21], we proposed a DC formulation of a clustering model with weights on variables and designed an appropriate DCA scheme for it. Numerical results on real word datasets are promising. Improvements of the algorithm for massive datasets are in progress.

7. Collaborations and Contracts

7.1. Contracts with Industry

7.2. National Initiatives

7.3. International Initiatives

Hoai An Le Thi is the project leader of InnoMaD Innovations techniques d'optimisation pour le traitement Massif de Données, FEDER (Fonds Européens de Développement Régional) project 2009-2012, Total project cost: 355 263 euros, FEDER grant: 159 868 euros.

Hoai An Le Thi is

- Distinguished visitor professor of University of Auckland October-November 2012 (3 weeks),

- Visiting professor of University Technology of Sydney (one week) in November 2012

- Member of Council of Department ICT, University of Science and Technology of Hanoi (USTH). USTH is a new university model in Viet nam, funded and developed in partnership with France, through a consortium of french higher education and research institutions (among them the university of Lorraine).

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) and the International Journal On Advances in Software. She has been a member of the program committee of the following conferences: DBKDA, eKNOW, AFDC, and SITIS and reviewer for TNNLS and RNTI. She is a member of 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'12) and the "Conférence Francophone sur l'Apprentissage Automatique" (CAp'12). He is an expert for the ANR.
- Hoai An Le Thi is a member of Editorial Board of 5 journals: Transactions on Computational Collective Intelligence (Springer), Journal of Optimization: Theory, Methods and Applications (GIP), Journal of Advanced Research in Computer Science (IASR), International Journal of Economics and Management Engineering, Vietnam Journal of Computer Science (a new journal in Springer). She is the editor of the special issue "Optimization and Learning", International Journal of Intelligent Information and Database Systems (IJIIDS), 2012.

  Hoai An Le Thi is

  - Member of Steering committee of the annual international conferences ACIIDS
  - Co-chair of Steering committee of the annual International Conference on Computer Science, Applied Mathematics and applications
  - Member of Scientific Committee ICCCI international conference on Computational Collective Intelligence (ICCCI), Ho Chi Minh city, 28-30 November, 2012
  - Organizer of Modelling and Optimization Techniques for Business Intelligence MOTBI'2012, workshop in 5th international conference on Computational Collective Intelligence (ICCCI), Ho Chi Minh city, 28-30 November, 2012.

8.2. Committees memberships

- Yann Guermeur was a member ("rapporteur") of the PhD jury of Juliana Silva Bernardes (Université Paris 6 and Universidade Federal do Rio de Janeiro).

8.3. Vulgarization

8.4. Invited Conferences

Hoai An Le Thi is invited speaker of the International Scientific School on Knowledge Management 2012, Quang Binh University (Viet nam), 25-26 Nov. 2012.

8.5. Teaching

- Rémi Bonidal gave lectures at the UL.

- Martine Cadot is PRAG in the Department of Computer Science at the UL 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 UL where he teaches machine learning to master (M1) students.
- Fabienne Thomarat is Associate Professor at the École Nationale Supérieure des Mines de Nancy / UL (engineering school, master of engineering school). She is in charge of the option bioinformatics at the Department of Computer Science.

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] F. Lauer and Y. Guermeur. MSVMpack: a multi-class support vector machine package. *Journal of Machine Learning Research*, 12:2293–2296, 2011.

[7] 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

[8] F. Abdat, M. Amouroux, Y. Guermeur, W. Blondel, “Hybrid feature selection and SVM-based classification for mouse skin precancerous stages diagnosis from bimodal spectroscopy”, *Optics Express 20*, 1, 2012, p. 228–244.

[9] R. Bonidal, S. Tindel, Y. Guermeur, “Model selection for the ℓ_2-SVM by following the regularization path”, (in revision).

[10] M. Cadot, A. Lelu, “Combining Explicitness and Classifying Performance via MIDOVA Lossless Representation for Qualitative Datasets”, *International Journal On Advances in Software 5*, 1&2, 2012, p. 1–16.

[11] S. Cheng, H. Le Thi, “Learning sparse classifiers with Difference of Convex functions Algorithms”, *Optimization Methods and Software*, (in press).

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

[13] Y. GUERMEUR, "Combining Multi-Class SVMs with Linear Ensemble Methods that Estimate the Class Posterior Probabilities", *Communications in Statistics - Theory and Methods*, 2013, (in press).

[14] F. LAUER, "Estimating the probability of success of a simple algorithm for switched linear regression", *Nonlinear Analysis: Hybrid Systems 8*, 2013, p. 31–47, Supplementary material available at `http://www.loria.fr/~lauer/klinreg/`.

[15] H. LE, H. LE THI, D. PHAM, V. HUYNH, "Block Clustering based on DC programming and DCA", *Neural Computation*, 2013.

International Peer-Reviewed Conference/Proceedings

[16] R. BONIDAL, F. THOMARAT, Y. GUERMEUR, "Estimating the class posterior probabilities in biological sequence segmentation", *in : SMTDA'12*, 2012.

[17] J. BUSSET, M. CADOT, "Démêler les actions des articulateurs en jeu lors de la production de parole avec le logiciel C.H.I.C. : analyse de séquences de radiographies de la tête", *in : ASI6*, p. 284–298, 2012.

[18] J. BUSSET, M. CADOT, "Fouille d'images animées : cinéradiographies d'un locuteur", *in : FOSTA'13*, p. 1–12, 2013. (accepted).

[19] F. LAUER, V. LE, G. BLOCH, "Learning smooth models of nonsmooth functions via convex optimization", *in : MLSP'12*, 2012.

[20] A. LE, H. LE THI, M. NGUYEN, A. ZIDNA, "Network Intrusion Detection Based on Multi-Class Support Vector Machine", *in : Computer and Computational Intelligence*, p. 536–543, 2012.

[21] H. LE, M. TA, H. LE THI, D. PHAM, "DC Programming and DCA for clustering using weighted dissimilarity measures", *in : 5th NIPS Workshop on Optimization for Machine Learning*, 2012.

[22] H. LE THI, B. NGUYEN, H. LE, "Sparse signal recovery by Difference of Convex functions Algorithms", *in : Intelligent Information and Database Systems*, (to appear).

[23] A. LELU, M. CADOT, "A Proposition for Fixing the Dimensionality of a Laplacian Low-rank Approximation of any Binary Data-matrix", *in : eKnow'13*, 2013. (accepted).

[24] L. MAXIM, M. CADOT, P. MANSIER, "Should scientists communicate uncertainty to the public in health controversies? The case of endocrine disrupters' effects on male fertility", *in : Between Scientists & Citizens*, p. 263–274, 2012.

[25] F. THOMARAT, F. LAUER, Y. GUERMEUR, "Cascading Discriminant and Generative Models for Protein Secondary Structure Prediction", *in : PRIB'12*, p. 166–177, 2012.

Scientific Books (or Scientific Book chapters)

[26] A. LELU, M. CADOT, "Détecter les ruptures thématiques dans les discours : synergie entre supervision et non-supervision", *in : DEFT*, Hermès, 2012, p. 49–63.

References in notes

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

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

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

[30] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. pages 1–34, 1996.

[31] Y. Guermeur. Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, 5(2):168–179, 2002.

[32] 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).

[33] H.A. Le Thi. DC programming and DCA. `http://lita.sciences.univ-metz.fr/~lethi/DCA.html`.

[34] H.A. Le Thi and D.T. Pham. The DC (Difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133:23–46, 2005.

[35] D.T. Pham and H.A. Le Thi. Optimization algorithms for solving the trust region subproblem. *SIAMJ. Optimization*, 8(2):476–505, 1998.

[36] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA, 2002.

[37] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. V.H. Winston & Sons, Washington, D.C., 1977.

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

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