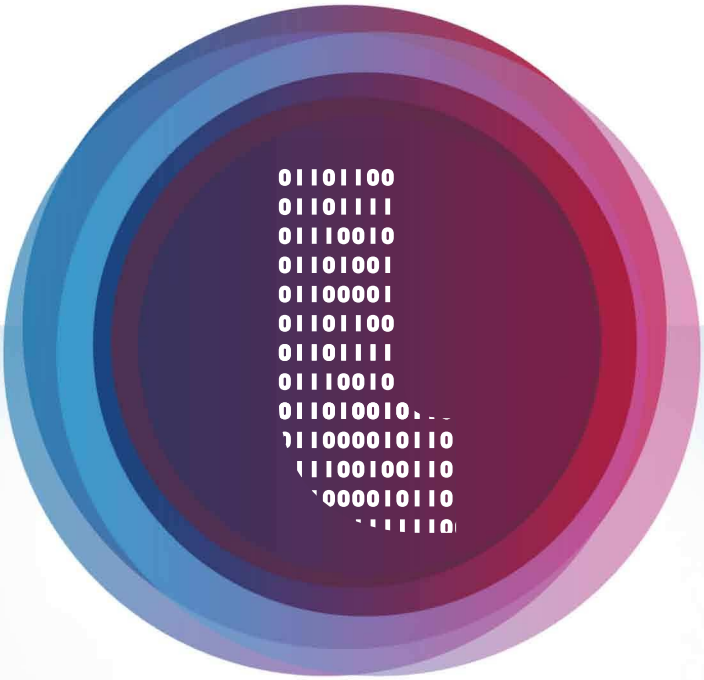
Loria
Laboratoire lorrain de recherche en informatique et ses applications

Lorraine Laboratory of
Research in Computer Science
and its Applications

# ACTIVITY REPORT 2013



TEAM

## ABC

## Apprentissage et Biologie Computationnelle

Research / Training / Transfer
in an international context

# Table of contents


**1. Team** . . . . . **1**

**2. Overall Objectives** . . . . . **1**

2.1. Highlights of the year 1

**3. Scientific Foundations** . . . . . **2**

3.1. Introduction 2

3.2. Statistical learning theory 2

3.3. Robust data mining and classification 2

3.4. Optimization in Machine Learning and Data Mining (MLDM) 3

**4. Application Domains** . . . . . **3**

4.1. Molecular biology 3

4.2. Image analysis 4

**5. Software** . . . . . **4**

5.1. MSVMpack 4

**6. New Results** . . . . . **4**

6.1. Statistical learning theory 4

6.1.1. Ensemble methods for M-SVMs 4

6.1.2. Guaranteed risks for large margin multi-category classifiers 5

6.2. Biological sequence segmentation 5

6.3. Robust data mining and classification 5

6.3.1. Robust data mining and modeling for complex data 5

6.3.2. Development of a method for dimensionality reduction 6

6.4. Regression and optimization 6

6.4.1. Geometric approach to switching linear regression 6

6.4.2. DC programming for switching linear regression 6

6.4.3. Switching regression via sparse optimization and support vector machines 7

6.4.4. Iteratively reweighting scheme for sparse recovery 7

6.5. Nonconvex optimization techniques for MLDM 7

**7. Dissemination** . . . . . **9**

7.1. Scientific Animation 9

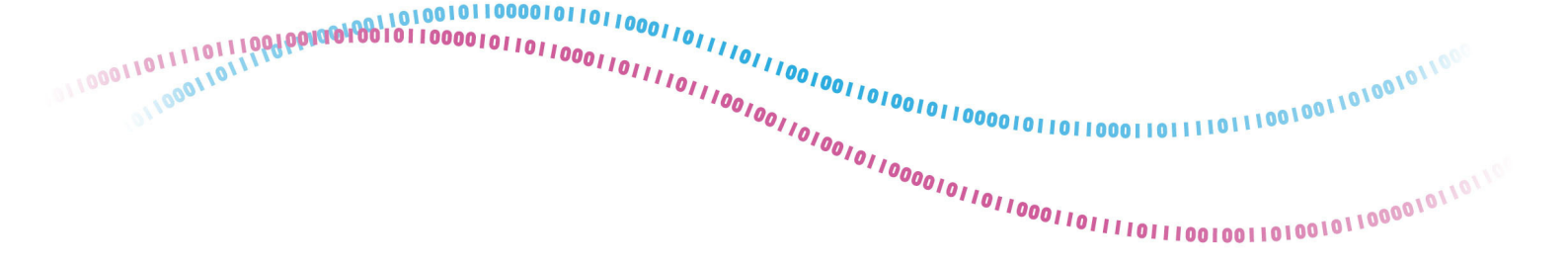
7.2. Committees memberships 9

7.3. Vulgarization 9

7.4. Invited Conferences 9

7.5. Teaching 9

**8. Bibliography** . . . . . **9**

# 1. Team

**Research scientists**

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, until August 2013]
Fabienne Thomarat [Associate Professor, UL, until October 2013]

**Administrative staff**

Aurélie Adam-Defeux [Assistant, CNRS]

**Technical staff**

Emmanuel Didiot [Research Engineer, CNRS, since September 2013]

**PhD students**

Rémi Bonidal [UL; defended in June 2013]
Hafida Bouziane-Chouarfia [Assistant Professor, USTO; defense planned in 2014]
Mounia Hendel [Assistant Professor, USTO]
Edouard Klein [UL; defended in November 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) [3]. Our applications are in the field of biological sequence processing. Precisely, our research themes can be summarized as follows:

- Derivation of bounds on the risk of classifiers
- Development of methods of model selection
- Specification, implementation, and evaluation of multi-class support vector machines
- Specification, implementation, and evaluation of kernel methods for regression
- Robust data mining and classification
- 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. They are roughly uniformly distributed between these three fields.

## 2.1. Highlights of the year

The main highlight for year 2013 is the start of the "Apprentissage statistique pour la segmentation de séquences biologiques" (A3SB) project funded by the CNRS.

# 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 [44] 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 [42]. In that framework, the objective function of the training algorithm incorporates a term related to the *capacity* of $\mathscr{F}$ [37]. The most famous example of implementation of this principle is provided by the training algorithms of the SVMs. Those machines are *kernel machines* [41] conceived to compute indicator functions, to perform regression or to estimate densities. They have been introduced by Vapnik and his co-workers [31, 33], as nonlinear extensions of the maximal margin hyperplane [43]. Their main advantage is that they can avoid overfitting in the case where the size of the sample is small [44, 36]. Several M-SVMs have been proposed in literature (see [3] for a survey). There theory can be embeded in the theory of large margin multi-category classifiers [35, 2].

## 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 [34]: texts, biological and speech sequences, social and behavioural data. Robust Data Mining consists, to our eyes [32], 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) ([40, 39, 38] 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, $\phi$-learning, clustering, ... see the incomplete list of references in [38]). 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, 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: $\alpha$*-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.

## 4.2. Image analysis

**Participants**: Hoai An Le Thi.

Compressed Sensing or Compressive Sensing (CS) is an emerging area having significant 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. MSVMpack

**Participants**: Emmanuel Didiot, Fabien Lauer [contact].

This package implementing all the M-SVMs published so far [5, 3] 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 3000. Updates are regularly released.

Since September 2013, we are working on a Windows port of MSVMpack to further increase its popularity, especially in the community of biologists.

# 6. New Results

## 6.1. Statistical learning theory

### 6.1.1. Ensemble methods for M-SVMs

**Participants**: Yann Guermeur.

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 the 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 each model has its advantages and drawbacks. These observations call for the evaluation of combinations of M-SVMs. In [9],

we studied the combination of M-SVMs with linear ensemble methods (LEMs). Their sample complexity is low, which prevents them from overfitting, and the outputs of two of them are estimates of the class posterior probabilities.

### 6.1.2. Guaranteed risks for large margin multi-category classifiers

**Participants**: Yann Guermeur.

In 2007, we contributed to the Vapnik-Chervonenkis theory of large margin multi-category classifiers by introducing the appropriate class of generalized Vapnik-Chervonenkis dimensions: the class of $\gamma$-$\Psi$-dimensions. The guaranteed risk we derived back then exhibited a suboptimal $\frac{\ln(m)}{\sqrt{m}}$ convergence rate. In 2013, we established a sharper bound, whose convergence rate is $\sqrt{\frac{\ln(m)}{m}}$.

## 6.2. Biological sequence segmentation

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

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 [30], we introduce the latest version of the application. It is evaluated on one specific task: protein secondary structure prediction.

## 6.3. Robust data mining and classification

### 6.3.1. Robust data mining and modeling for complex data

**Participants**: Martine Cadot.

Since her thesis [32], Martine Cadot develops the MIDOVA method for expressing the structure of a Variables X Individuals table as the set of all the "itemsets" of variables necessary and sufficient for reconstructing the data. These itemsets embed the any level-interactions between the variables, whatever their lengths. 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 a number of data mining tasks, such as supervised learning, which was used to validate the method. 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.

Work directions are to be carried on in order to achieve and disseminate this method as a software tool for data miners:

- direction 1: a line of thought on extracting models from data [29] was initiated by analyzing how the ASI (Analyse Statistique Implicative) produces a set of rules for causal reasoning.

- direction 2: Hitherto confined to binary variables, we need to go for dynamic quantitative and qualitative variables. We used a 3-way data factorization method for cineradiographic data [18] which highlighted the importance of taking into account the relationships of level> 2 between the variables, and the statistical validation of these relationships. To extract knowledge from complex data, usual data mining has to be combined indeed with inferential statistics.

Remains to be done:

- Consolidating directions 1 and 2 and bring MIDOVA to them.

- Developing direction 3 : setting up an efficient software for the method by parallelizing the algorithms.

### 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 [26], 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.

## 6.4. Regression and optimization

### 6.4.1. Geometric approach to switching linear regression

**Participants**: Fabien Lauer.

In [24], we considered switching linear regression problems, i.e., regression problems where the data are assumed to be generated by a model switching between multiple linear submodels. We proposed a new approach based on the geometric properties of switching regression models in parameter space. More precisely, the data are mapped in that space such that each submodel is represented by a hypersphere. Then, we showed how these hyperspheres can be easily separated by Principal Component Analysis (PCA) and derived a condition under which this separation is optimal for models with two submodels. From this classification, classical (robust) regression can be applied independently to each group of data to estimate the model parameters from the classified data set. A simple procedure was also proposed to extend the method to the case of a number of submodels greater than two. Experiments showed that the final algorithm can accurately estimate both the parameters and the number of submodels while being simple to apply and far more robust to noise than other methods.

### 6.4.2. DC programming for switching linear regression

**Participants**: Hoai An Le Thi, Fabien Lauer.

In [16], 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.

### 6.4.3. Switching regression via sparse optimization and support vector machines

**Participants**: Fabien Lauer.

In [23], we extended the recent sparse optimization approach for switching regression to the case of nonlinear submodels belonging to a reproducing kernel Hilbert space, while focusing on the connections with support vector machines. This approach is based on a convex relaxation of a sparse optimization problem, where the submodels are iteratively estimated one by one by maximizing the sparsity of the corresponding error vector. We modified this approach in several ways. First, we relaxed the sparsity condition ensuring success by introducing robust sparsity, which can be optimized through the minimization of a modified $\ell_1$-norm or, equivalently, of the $\varepsilon$-insensitive loss function. Then, we showed that, depending on the choice of regularizer, the method is equivalent to different forms of support vector regression. More precisely, the submodels can be estimated by iteratively solving a classical support vector regression training problem, in which the sparsity of support vectors relates to the sparsity of the error vector in the considered switching regression framework. This allowed us to extend theoretical results as well as efficient optimization algorithms from the field of support vector machines to this framework.

The notion of robust sparsity was also used in [17] to estimate switching linear multivariate models, with a particular focus on the application to the identification of multiple-intput-multiple-output hybrid dynamical systems in state-space form.

### 6.4.4. Iteratively reweighting scheme for sparse recovery

**Participants**: Fabien Lauer.

Motivated by the sparse optimization approach to switching regression described above, we investigated in [11] 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 $\ell_1$-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.

## 6.5. Nonconvex optimization techniques for MLDM

**Participants**: Hoai An Le Thi.

1. New and efficient DCA based algorithms for Minimum Sum-of-Squares Clustering [14]

We develop new efficient approaches based on DC (Difference of Convex functions) programming and DCA (DC Algorithm) to perform clustering via Minimum Sum-of-Squares Euclidean distance. We consider the two most widely used models for the so called Minimum Sum-of-Squares Clustering (MSSC in short) that are a bilevel programming problem and a mixed integer program. Firstly, the mixed integer formulation of MSSC is carefully studied and is reformulated as a continuous optimization problem via a new result on exact penalty technique in DC programming. DCA is then investigated to the resulting problem. Secondly, we introduce a Gaussian Kernel version of the bilevel programming formulation of MSSC, named GKMSSC. The GKMSSC problem is formulated as a DC program for which a simple and efficient DCA scheme is developed.

A regularization technique is investigated for exploiting the nice effect of DC decomposition and a simple procedure for finding good starting points of DCA is developed. The proposed DCA schemes are original and very inexpensive because they amount to computing, at each iteration, the projection of points onto a simplex and/or onto a ball, and/or onto a box, that are all determined in the explicit form. Numerical results on real word datasets show the efficiency, the scalability of DCA and its great superiority with respect to k-means and Kernel k-means, standard methods for clustering.

2. Block Clustering based on Difference of Convex functions (DC) Programming and DC Algorithms [12]

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 such as Two-mode K-means, Two-mode Fuzzy Clustering, Block Classification EM.

3. Binary classification via spherical separator by DC programming and DCA [13]

We consider a binary supervised classification problem, called spherical separation, that consists of finding, in the input space or in the feature space, a minimal volume sphere separating the set A from the set B (i.e., a sphere enclosing all points of A and no points of B). The problem can be cast into the DC (Difference of Convex functions) programming framework and solved by DCA (DC Algorithm) as shown in the works of Astorino et al (J. Glob Optim 48(4):657-669, 2010). The aim of this paper is to investigate more attractive DCA based algorithms for this problem. We consider a new optimization model and propose two interesting DCA schemes. In the first scheme we have to solve a quadratic program at each iteration, while in the second one all calculations are explicit. Numerical simulations show the efficiency of our customized DCA with respect to the methods developed in Astorino et al.

4. Learning sparse classifiers with Difference of Convex functions Algorithms [8]

Sparsity of a classifier is a desirable condition for high dimensional data and large sample sizes. This paper investigates 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 $\ell_2$, $\ell_1$, approximate $\ell_0$, and capped $\ell_1$ regularization. We propose two new objective functions that further promote sparsity, corresponding to the ramp loss versions of approximate $\ell_0$ and capped $\ell_1$ 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 $\ell_1$ 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.

5. Feature selction in S3VM [22] and M-SVM [19]

We develop an efficient feature selection method using zero-norm $l_0$ in the context of Semi-Supervised Support Vector Machine (S3VM) [22] and Multiple Support Vector Machine (M-SVM) [19] Using appropriate continuous approximations of $l_0 - norm$, we reformulate the problem as DC (Difference of Convex functions) programs. DCA (DC Algorithm) is then developed to solve the resulting problem. Computational experiments on several real-world datasets show the efficiency and the scalability of our method.

6. Learning with uncertain data by robust optimization using DC programming and DCA [20]

We consider the problem of feature selection and classification on uncertain data that is inherently prevalent in almost all datasets. Using principles of Robust Optimization, we propose robust schemes to handle data with ellipsoidal model and box model of uncertainty. The difficulty in treating $\ell_0$-norm in feature selection problem is overcome by using appropriate approximations and Difference of Convex functions (DC) programming and DC Algorithms (DCA). The computational results show that the proposed robust optimization approaches are superior than a traditional approach in immunizing perturbation of the data.

# 7. Dissemination

## 7.1. 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, 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 "Conférence Francophone sur l'Apprentissage Automatique" (CAp'13) and the "6èmes Journées de la société française de chémoinformatique" (SFCi'13). He is an expert for the ANR.

## 7.2. Committees memberships

Yann Guermeur was a member ("rapporteur") of the PhD juries of Mohammed Hindawi (INSA de Lyon) and Thanh-Nghi Doan (Université de Rennes 1).

## 7.3. Vulgarization

## 7.4. Invited Conferences

Le Thi Hoai An was an invited speaker at EURO-INFORMS Joint International Meeting on Operational Research, Rome July 1-4, 2013. She gave a talk entitled "Difference of convex functions optimization".

## 7.5. Teaching

- Martine Cadot is PRAG in the Department of Computer Science at the UL where she teaches statistics and 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.

# 8. 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. 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.

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

[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] R. BONIDAL, S. TINDEL, Y. GUERMEUR, "Model selection for the $\ell_2$-SVM by following the regularization path", *Transactions on Computational Collective Intelligence Vol. XIII (LNCS 8342)*, 2014, p. 83–112.

[8] CHENG SOON ONG, LE THI HOAI AN, "Learning sparse classifiers with difference of convex functions algorithm", *Optimization Methods and Softwar 28*, 4, 2013, p. 830–854.

[9] Y. GUERMEUR, "Combining Multi-class SVMs with Linear Ensemble Methods that Estimate the Class Posterior Probabilities", *Communications in Statistics - Theory and Methods 42*, 16, 2013, p. 2311–2330.

[10] 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/`.

[11] V. L. LE, F. LAUER, G. BLOCH, "Selective $\ell_1$ minimization for sparse recovery", *IEEE Transactions on Automatic Control*, 2013, (to appear).

[12] LE HOAI MINH, LE THI HOAI AN, PHAM DINH TAO, HUYNH VAN NGAI, "Block Clustering based on DC programming and DCA", *NECO Neural Computation 25*, 10, 2013, p. 2776–2807.

[13] LE THI HOAI AN, LE HOAI MINH, PHAM DINH TAO, NGAI VAN HUYNH, "Binary classification via spherical separator by DC programming and DCA", *J. Global Optimization 56*, 4, 2013, p. 1393–1407.

[14] LE THI HOAI AN, LE HOAI MINH, PHAM DINH TAO, "New and efficient DCA based algorithms for Minimum Sum-of-Squares Clustering", *Pattern Recognition 47*, 1, 2014, p. 388–401.

[15] LE THI HOAI AN, LE MINH TAM, NGUYEN BICH THUY, "A novel approach to automated cell counting based on DCA", *Computational Collective Intelligence*, 2013, (to appear).

[16] 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*, 2014, (to appear).

## International Peer-Reviewed Conference/Proceedings

[17] L. BAKO, V. L. LE, F. LAUER, G. BLOCH, "Identification of MIMO switched state-space models", *in : Proc. of the American Control Conference (ACC), Washington, DC, USA*, 2013.

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

[19] HOAI AN LE THI, MANH CUONG NGUYEN, "Efficient algorithms for Feature Selection in Multi-class Support Vector Machine", *in : Advanced Computational Methods for Knowledge Engineering*, 2013.

[20] HOAI AN LE THI, VO XUAN THANH, PHAM DINH TAO, "Robust Feature Selection for SVMs under Uncertain Data", *in : MLDM*, p. 151–165, 2013.

[21] HOAI MINH LE, BICH THUY NGUYEN THI, MINH THUY TA, HOAI AN LE THI, "Segmentation via Feature Weighted Fuzzy Clustering by a DCA based algo", *in : Advanced Computational Methods for Knowledge Engineering*, 2013.

[22] HOAI MINH LE, HOAI AN LE THI, MANH CUONG NGUYEN, "DCA based algorithms for feature selection in Semi-Supervised Support Vector Machines", *in : MLDM*, 2013.

[23] V. L. LE, F. LAUER, L. BAKO, G. BLOCH, "Learning nonlinear hybrid systems: from sparse optimization to support vector regression", *in : Proc. of the 16th ACM Int. Conf. on Hybrid Systems: Computation and Control (HSCC), Philadelphia, PA, USA*, p. 33–42, 2013.

[24] V. L. LE, F. LAUER, G. BLOCH, "Identification of linear hybrid systems: a geometric approach", *in : Proc. of the American Control Conference (ACC), Washington, DC, USA*, 2013.

[25] LE THI HOAI AN, BICH THUY NGUYEN THI, LE HOAI MINH, "Sparse Signal Recovery by Difference of Convex Functions Algorithms", *in : ACIIDS*, p. 387–397, 2013.

[26] A. LELU, M. CADOT, "A proposition for fixing the dimensionality of a Laplacian low-rank approximation of any binary data-matrix", *in : eKNOW'13*, p. 70–73, 2013.

[27] MINH THUY TA, HOAI AN LE THI, LYDIA BOUDJELOUD-ASSALA, "Clustering data streams over sliding windows by DCA", *in : Advanced Computational Methods for Knowledge Engineering*, 2013.

[28] MINH THUY TA, LE THI HOAI AN, LYDIA BOUDJELOUD-ASSALA, "An Efficient Clustering Method for Massive Dataset Based on DC Programming and DCA", *in : ICONIP*, p. 538–545, 2013.

### Scientific Books (or Scientific Book chapters)

[29] M. CADOT, "Modèle des données à base de règles : de la construction au pilotage", *in : L'analyse statistique implicative - Méthode exploratoire et confirmatoire à la recherche de causalités - 2e édition*, R. Gras, J.-C. Régnier, C. Marinica, and F. Guillet (editors), Cépaduès, 2013, p. 299–312.

[30] 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).

## References in notes

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

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

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

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

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

[36] 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).

[37] 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.

[38] H.A. LE THI. DC programming and DCA. `http://lita.sciences.univ-metz.fr/~lethi/DCA.html`.

[39] 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.

[40] D.T. PHAM AND H.A. LE THI. Optimization algorithms for solving the trust region subproblem. *SIAMJ. Optimization*, 8(2):476–505, 1998.

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

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

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

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