INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team ABC

Apprentissage et Biologie Computationnelle

Lorraine

THEME BIO

Activity Report

2007

Table of contents

1. Team **1**

2. Overall Objectives **1**

2.1. Introduction 1

2.2. Research themes 1

2.3. Scientific and industrial relations 1

2.4. Highlights of the year 1

3. Scientific Foundations **2**

3.1. Introduction 2

3.2. Statistical learning 2

3.3. Classification 2

4. Application Domains **2**

5. Software **3**

6. New Results **4**

6.1. Bounds on the risk of large margin multi-category classifiers 4

6.2. Model selection for multi-class SVMs 4

6.3. Protein secondary structure prediction 4

6.4. Molecular phylogeny 5

6.5. Classification 5

7. Other Grants and Activities **5**

7.1. Regional projects 5

7.2. National projects 6

8. Dissemination **6**

8.1. Serving the scientific community 6

8.2. Teaching 6

9. Bibliography **6**

1. Team

Head of team
Yann Guermeur [Research Associate (CR) CNRS, HdR]

Administrative assistants
Zohra Marzak [Secretary (SAR) INRIA, until 11/2007]
Sandra Bezon [Secretary (SAR) INRIA, since 11/2007]

Permanent researchers
Fabienne Thomarat [Associate Professor, INPL]

Post-doctoral fellows
Emmanuel Didiot [ATER, Nancy 2, since 09/2007]
Emmanuel Monfrini [UHP, until 03/2007]

Technical staff
Julien Vannesson [IE, from 09/2007 until 11/2007]

Internships
Julien Vannesson [ESIAL, from 03/2007 until 08/2007]
Aurélie Colas [ENSMN, from 03/2007 until 08/2007]

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

2. Overall Objectives

2.1. Introduction

The aim of the ABC 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 SVMs (M-SVMs). Our applications are in the field of biological sequence processing.

2.2. Research themes

- Bounds on the risk of classifiers
- Model selection
- Protein secondary structure prediction
- Molecular phylogeny

2.3. Scientific and industrial relations

- Participation in the "Génopole Strasbourg Alsace-Lorraine"
- Participation in the Bioinformatics project of the Région Lorraine
- Participation in the "Décrypthon" programme
- Various national and international collaborations
  - Laboratoire "Maturation des ARN et Enzymologie Moléculaire" (MAEM), UMR 7567, Nancy
  - Institut de Biologie et Chimie des Protéines, IBCP, Lyon
  - DFG Research Center Matheon, Berlin, Germany
  - Institute for Genomics and Bioinformatics, University of California, Irvine, USA

2.4. Highlights of the year

This year saw the first defence of an "Habilitation à Diriger des Recherches" by a member of the team [10].

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 occurs when 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, occurs when the set of categories is unknown.

3.2. Statistical learning

Statistical learning theory [20] is one of the fields of inferential statistics the bases of which have been established by V.N. Vapnik in the late 1960s. 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 on pairs made up of observations and labels, and a class of functions, of cardinality ordinarily infinite, the goal is to find in the class a function with optimal performance. Training can thus be reformulated as an optimization problem. In many cases, the objective function is related to the capacity of the class of functions [8]. The learning tasks considered belong to one of the three following areas: pattern recognition (discriminant analysis), function approximation (regression) and density estimation.

This theory considers more specifically two inductive principles. The first one, named empirical risk minimization (ERM) principle, consists in minimizing the training error. If the sample is small, one substitutes to this the structural risk minimization (SRM) principle. It consists in minimizing an upper bound on the expected risk (generalization error), a bound sometimes called a guaranteed risk. This latter principle is implemented in the training algorithms of the support vector machines (SVMs), which currently constitute the state-of-the-art for numerous problems of pattern recognition.

SVMs are connectionist models conceived to compute indicator functions, to perform regression or to estimate densities. They have been introduced during the last decade by Vapnik and co-workers [16], as nonlinear extensions of the maximal margin hyperplane [19]. Their main advantage is that they can avoid overfitting in the case where the size of the sample is small [20], [15].

3.3. Classification

Classification consists in clustering a set of objects in a finite number of categories unknown *a priori*. Objects are represented by their descriptions, which usually correspond to vectors of $\mathbb{R}^n$. Different types of methods can be used, gathered by literature into three main groups: mixture models, partitional clusterings and hierarchical clusterings. The objects can thus be clustered in various structures. While some methods only provide a partition of the objects, others generate hierarchical structures. We are primarily interested in hierarchical classifications, when the categories are arranged in a tree-structure.

Molecular phylogeny aims at identifying the evolutionary relationships between species, relationships which are represented by a tree. The descriptors of the species are usually derived from the sequence of a gene or a protein. To build a tree, different kinds of methods are used. Some of them are based on the characters (nucleotides or amino acids), the most common ones being the *maximum parsimony* method and the *maximum likelihood* method. The other methods are based on distances. Using a given criterion, the distance matrix methods aim at building a tree from a matrix of pairwise distances.

4. Application Domains

4.1. Molecular biology

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

Participants: Alexander Bockmayr, Emmanuel Didiot, Yann Guermeur, Emmanuel Monfrini, Fabienne Thomarat, Julien Vannesson.

Molecular biology is concerned with the study of three types of biological macromolecules: DNA, RNA, and proteins. Each of these molecules can initially be viewed as a string on a finite alphabet: DNA and RNA are nucleic acids made up of nucleotides A,C,G,T and A,C,G,U, respectively. Proteins are sequences of amino acids, which may be represented by an alphabet of 20 letters.

Molecular biology studies the information flow from DNA to RNA, and from RNA to proteins. In a first step, called *transcription*, a DNA string ("gene") is transcribed into messenger RNA (mRNA). In the second step, called *translation*, the mRNA is translated into a protein: each triplet of nucleotides encodes one amino acid according to the genetic code. The genes of eukaryotic cells are mostly composed of a succession of coding regions, called exons, and non-coding regions, called introns. During transcription, an intermediate step, the *splicing* process, is then necessary to remove the introns from the premessenger RNA. The remaining exons are concatenated yielding the mature RNA molecule. *Alternative splicing* is a regulatory mechanism by which variations in the incorporation of the exons into mRNA leads to the production of different forms of mature mRNAs and consequently to more than one related protein, or isoform.

Biological macromolecules are not just sequences of nucleotides or amino acids. Actually, they are complex three-dimensional objects. DNA shows the famous double-helix structure. RNA and proteins fold into complex three-dimensional structures, which depend on the underlying sequence. RNA is a single-stranded chain of nucleotides. However, a nucleotide in one part of the molecule can base-pair with a nucleotide in another part, following the Watson-Crick complementarity rules. This results in a folding of the molecule. The *secondary structure* of RNA indicates the set of base pairings in the three dimensional structure of the molecule. This information can be represented by a graph.

Proteins have several levels of structure. Above the primary structure (i.e., the sequence) is the *secondary structure*, which involves three basic types: α*-helices*, β*-sheets*, and aperiodic structure elements called *loops*. The spatial relationship of the secondary structures forms the tertiary structure. Several proteins can function together in a protein complex whose structure is referred to as the quaternary structure. A *domain* of a protein is a combination of secondary structure elements with some specific function. It contains an *active site* where an interaction with an external molecule may happen. A protein may have one or several domains.

The ultimate goal of molecular biology is to understand the *function* of biological macromolecules in the life of the cell. Function results from the *interaction* between different macromolecules, and depends on their structure. The overall challenge is to make the leap from sequence to function, through structure: the prediction of structure will help to predict the function.

Thanks to the huge number of gene and protein sequences available in the sequence databases, molecular phylogenetic analyses multiplied since a few decades. Molecular phylogeny [17] is the use of genes or protein sequences to gain information on the evolutionary history of organisms. By comparison of the sequence of a gene in different organisms, the evolutionary history of these sequences can be inferred. Based on the hypothesis that these sequences are orthologs (i.e., come from a same ancestral sequence by speciation events), the evolutionary history of the organisms can also be inferred and be represented by a tree.

5. Software

5.1. M-SVM: Multi-class Support Vector Machine

Participants: Yann Guermeur [correspondent], Julien Vannesson.

We have extended the functionalities of our application dedicated to protein sequence processing [7]. This application can now process multiple alignments [14]. Adding this functionality required to develop several auxiliary tools, to clean up alignments and prepare them so that they could be processed by the M-SVM.

6. New Results

6.1. Bounds on the risk of large margin multi-category classifiers

Keywords: *Statistical learning theory*, *generalized VC dimensions*, *support vector machines*.

Participant: Yann Guermeur.

In the context of discriminant analysis, Vapnik's statistical learning theory has mainly been developed in three directions: the computation of dichotomies with binary-valued functions, the computation of dichotomies with real-valued functions, and the computation of polytomies with functions taking their values in finite sets, typically the set of categories itself. The case of classes of vector-valued functions used to compute polytomies has seldom been considered independently, which is unsatisfactory, for three main reasons. First, this case encompasses the other ones. Second, it cannot be treated appropriately through a naïve extension of the results devoted to the computation of dichotomies. Third, most of the classification problems met in practice involve multiple categories.

In [11], [13], a Vapnik-Chervonenkis (VC) theory of large margin multi-category classifiers has been introduced. Central in this theory are generalized VC dimensions called the γ-Ψ-dimensions. First, we derived a uniform convergence bound on the risk of the classifiers of interest. The capacity measure involved in this bound is a covering number. This covering number can be upper bounded in terms of the γ-Ψ-dimensions thanks to generalizations of Sauer's lemma. We illustrated this property in the specific case of the scale-sensitive Natarajan dimension. A bound on this latter dimension was then computed for the class of functions on which multi-class SVMs are based. This makes it possible to apply the structural risk minimization inductive principle to those machines.

6.2. Model selection for multi-class SVMs

Keywords: *Multi-class support vector machines*, *leave-one-out error*, *radius-margin bounds*.

Participants: Yann Guermeur, Emmanuel Monfrini.

Using a support vector machine requires to set two types of hyperparameters: the soft margin parameter C and the parameters of the kernel. To perform this model selection task, one can use various procedures based on cross-validation. Obviously, the major drawback of such procedures rests in their time requirements. To overcome this difficulty, several upper bounds on the leave-one-out error of pattern recognition support vector machines have been derived. We have established a strict generalization of one of these bounds, called the radius-margin (RM) bound, to the case of the multi-class SVM of Weston and Watkins. In order to complete this result, we have extended the notion of "2-norm" (bi-class) SVM, by introducing a variant of the M-SVM of Lee, Lin and Wahba with quadratic loss, the M-SVM^2. As in the bi-class case, training the soft margin version of this machine appears equivalent to training a hard margin version of the same model, with a different kernel. This establishes the usefulness of the RM bound we have derived for the M-SVM of Lee, Lin and Wahba.

6.3. Protein secondary structure prediction

Keywords: *Protein secondary structure*, *clustering*, *multiple alignments*.

Participants: Yann Guermeur, Julien Vannesson.

It is well known that presenting multiple alignments instead of the sole primary structures in input of protein secondary structure prediction models drastically improves the prediction accuracy. We introduced this functionality in our prediction method based on a M-SVM [14]. Our study on the subject primarily dealt with the optimal way to combine the predictions made independently for the different sequences of an alignment. We also considered the application of various clusterings on the sequences of alignements used for training. Combining the best options resulted in an increase in the recognition rate which is statistically significant with confidence exceeding 0.95.

6.4. Molecular phylogeny

Keywords: *Molecular phylogeny*, *kernel methods*, *multiple alignments*.

Participant: Fabienne Thomarat.

One of the main advantages of kernels is the fact that they can be used to measure the similarity between objects which are not described by vectors of $\mathbb{R}^n$. They can, in particular, be used to measure the similarity between two sequences, similarity from which a distance can be derived. Pairwise sequence distances are broadly used in molecular phylogenetics. Their computation implies the use of statistical models of DNA/protein evolution. We investigate the benefits springing from substituting kernels to these distances.
The first step to infer the evolutionary history of gene or protein sequences consists in building an alignment of all these sequences, i.e., to determine the homology (common ancestry) at each site of the sequences. To that end, biologists usually use the algorithm provided by the ClustalW program [18]. The first step of this algorithm consists in the computation of a distance between each pair of sequences. These distances are then used to build a tree, referred to as the "guide tree". We have substituted to the initial method of distance calculation a simple kernel based on the k-uples shared by the two sequences. The few tests performed with it show that although the resulting guide trees are less plausible than the ones of reference, they contain some biologically relevant information. We go on testing kernels more adapted to the problem of interest.

6.5. Classification

Keywords: *Bayesian inference*, *MCMC simulation*, *Water meters*, *clustering*, *graphical models*, *number of clusters*.

Participants: Myriam Maumy, Frédéric Bertrand.

Water meters give a more and more inaccurate measure of water consumption, when getting older. Such a degradation is to give rise to an underestimation of consumption, which turns to be an issue for water distribution companies. As a result these companies designed management strategies of water meters in order to achieve two goals:

1. reduction of financial losses due to unaccounted-for water,
2. water metering equity between customers.

Every strategy needs as a prerequisite the understanding of the degradation process and the evaluation of the loss of accuracy. In [12], the ageing of water meters is described by a dynamic discrete state model. Each of these states characterizes a given measurement quality.

This model, together with the observation of measurement errors within each state, allows to estimate the ill-behaved water meters rate and the accuracy, as a function of the operating age of the device. Model parameters estimation and prediction of practically interesting quantities have been made by "Markov Chain Monte Carlo" (MCMC) simulation techniques.

The main aim of this study is to cluster the water meters of a water distribution company in several groups, whose number is to be determined, according to their estimated properties in order to be able to replace those that are likely to ill-behave soon and therefore complete the two aforementioned objectives.

7. Other Grants and Activities

7.1. Regional projects

We participate in the "Génopole Strasbourg Alsace-Lorraine" together with the laboratory "Maturation des ARN et Enzymologie Moléculaire" (MAEM) and the "Institut de Génétique et de Biologie Moléculaire et Cellulaire" (IGBMC) in Strasbourg.

7.2. National projects

Since October 2005, we participate in the project "Développement et utilisation d'approches informatiques et théoriques pour l'analyse des liens existant entre défauts d'épissage et maladies génétiques" funded by the Décrypthon programme: http://www.decrypthon.fr/. Our partner in this project is the MAEM laboratory.

8. Dissemination

8.1. Serving the scientific community

Yann Guermeur has been a member of the program committee of CAp'07 and the organizer of the special session "Supervised prediction with neural networks and SVMs" of ASMDA'07. He is an expert for the ANR.

8.2. Teaching

Fabienne Thomarat is Associate Professor at the École Nationale Supérieure des Mines de Nancy / Institut National Polytechnique de Lorraine (engineering school, master of engineering school). She is in charge of one option (Bioinformatics) at the Department of Computer Science.

Yann Guermeur has been teaching bioinformatics in the M2P speciality "Génomique et Informatique" of the Master "Sciences de la Vie et de la Santé" (SVS), at the University Henri Poincaré.

Emmanuel Didiot is ATER at the University Nancy 2.

9. Bibliography

Major publications by the team in recent years

[1] E. Balas, A. Bockmayr, N. Pisaruk, L. Wolsey. *On unions and dominants of polytopes*, in "Mathematical Programming, Ser. A", vol. 299, 2004, p. 223-239.

[2] A. Bockmayr, A. Courtois. *Using hybrid concurrent constraint programming to model dynamic biological systems*, in "18th International Conference on Logic Programming, ICLP'02, Copenhagen", Springer, LNCS 2401, 2002, p. 85-99.

[3] A. Bockmayr, J. N. Hooker. *Constraint Programming*, in "12: Discrete Optimization", K. Aardal, G. Nemhauser, R. Weismantel (editors), Handbooks in Operations Research and Management Science, chap. 10, Elsevier, 2005, p. 559–600.

[4] Y. Guermeur, A. Elisseeff, D. Zelus. *A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers*, in "Applied Stochastic Models in Business and Industry", vol. 21, n^{o} 2, 2005, p. 199–214.

[5] Y. Guermeur. *Combining discriminant models with new multi-class SVMs*, in "Pattern Analysis and Applications", vol. 5, n^{o} 2, 2002, p. 168–179.

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

[7] Y. GUERMEUR, A. LIFCHITZ, R. VERT. *A kernel for protein secondary structure prediction*, in "Kernel Methods in Computational Biology", B. SCHÖLKOPF, K. TSUDA, J.-P. VERT (editors), The MIT Press, 2004, p. 193–206.

[8] Y. GUERMEUR, O. TEYTAUD. *Estimation et contrôle des performances en généralisation des réseaux de neurones*, in "Apprentissage Connexionniste", Y. BENNANI (editor), chap. 10, Hermès, 2006, p. 279–342.

[9] V. Y. LUNIN, A. URZHUMTSEV, A. BOCKMAYR. *Direct phasing by binary integer programming*, in "Acta Crystallographica Section A", vol. 58, 2002, p. 283-291.

Year Publications

Doctoral dissertations and Habilitation theses

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

Articles in refereed journals and book chapters

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

Publications in Conferences and Workshops

[12] F. BERTRAND, M. MAUMY. *Application de méthodes de classification sur des vitesses métrologiques de dégradation de compteurs d'eau*, in "SFC'07", 2007.

[13] Y. GUERMEUR. *Scale-sensitive Ψ-dimensions: the capacity measures for classifiers taking values in $\mathbb{R}^Q$*, in "ASMDA'07", 2007.

Miscellaneous

[14] J. VANNESSON. *Contribution au développement d'une méthode hybride discriminante-générative de prédiction de la structure secondaire des protéines*, Technical report, Master informatique de Nancy, 2007.

References in notes

[15] C. BURGES. *A tutorial on support vector machines for pattern recognition*, in "Data Mining and Knowledge Discovery", vol. 2, n^{o} 2, June 1998, p. 121–167.

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

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

[18] J. THOMPSON, D. HIGGINS, T. GIBSON. *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*, in "Nucleic Acids Research", vol. 22, n^{o} 22, 1994, p. 4673–4680.

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

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