International Workshop on Current Challenges in Kernel Methods (CCKM06)

The official 2006 kernel workshop
"10 years of kernel machines"

Koninklijke Vlaamse Academie van België (KVAB)
Brussels, Belgium
27-28 November 2006

Workshop website
http://www.machine-learning.be/cckm06/

Sponsors
De Wetenschappelijke Onderzoeksgemeenschap (WOG)
"Machine Learning for Data Mining and its Applications"
De Koninklijke Vlaamse Academie van Belgie (KVAB)
The PASCAL network of excellence
PASCAL
www.pascal-network.org

Workshop organizers
Prof. Bernard Manderick, V.U.B.
Dr. Tijl De Bie, K.U.L.

Workshop programme

<table>
<tr><th></th><th>Monday 27 November</th><th>Tuesday 28 November</th></tr>
<tr><td>8.00-8.45</td><td>Coffee and registration</td><td rowspan="2">Coffee and registration</td></tr>
<tr><td>8.45-9.00</td><td>Opening</td></tr>
<tr><td>9.00-9.30</td><td rowspan="3">Jean-Philippe Vert</td><td rowspan="3">Ioannis Tsochantaridis</td></tr>
<tr><td>9.30-10.00</td></tr>
<tr><td>10.00-10.30</td></tr>
<tr><td>10.30-11.00</td><td>Coffee break</td><td>Coffee break</td></tr>
<tr><td>11.00-11.30</td><td rowspan="3">John Shawe-Taylor</td><td rowspan="2">Sandor Szedmak</td></tr>
<tr><td>11.30-12.00</td></tr>
<tr><td>12.00-12.30</td><td rowspan="5">Lunch break
and
Poster session</td></tr>
<tr><td>12.30-13.00</td><td>Andreas Christmann</td></tr>
<tr><td>13.00-13.30</td><td rowspan="3">Lunch break</td></tr>
<tr><td>13.30-14.00</td></tr>
<tr><td>14.00-14.30</td></tr>
<tr><td>14.30-15.00</td><td rowspan="2">Johan Suykens</td><td rowspan="2">Ingrid Daubechies</td></tr>
<tr><td>15.00-15.30</td></tr>
<tr><td>15.30-16.00</td><td rowspan="2">Kristiaan Pelckmans</td><td rowspan="2">Alain Rakotomamonjy</td></tr>
<tr><td>16.00-16.30</td></tr>
<tr><td>16.30-17.00</td><td>Coffee break</td><td>Coffee break</td></tr>
<tr><td>17.00-17.30</td><td rowspan="3">Bernhard Schoelkopf</td><td rowspan="2">Nello Cristianini</td></tr>
<tr><td>17.30-18.15</td></tr>
<tr><td>18.15-18.30</td><td>Closing</td></tr>
<tr><td>18.30-...</td><td>Reception with drinks and snacks
in the KVAB</td><td>Short guided tour of Brussels
Banquet in the heart of Brussels</td></tr>
</table>

Invited talks

Classification of biological sequences with kernel methods
Jean-Philippe Vert (Ecole des Mines de Paris)

The increasing amount of gene and protein sequences generated by sequencing projects has boosted the development of statistical and machine learning algorithms to compare and classify biological sequences. Kernel methods in general, and support vector machines in particular, lends themselves particularly well to this problem thanks to their capacity to manipulate any type of data as long as a positive definite kernel is defined. In this talk I will review different ways to make kernels for variable-length sequences and illustrate their applications in the context of biological sequence classification.

As this is the first talk in the workshop, I will start with a short introduction to kernel methods.

Kernel Methods: the Emergence of a Well-founded Machine Learning
John Shawe-Taylor (University College London)

During the last ten years kernel methods have established themselves as the method of choice in many types of machine learning problems. The talk will review the development of kernel methods emphasising a surprising convergence of theory and practice that we contend underscored the process. The talk will argue in favour of a broad definition of theory as a touchstone for the well-foundedness and scientific justification of an approach, but equally emphasise a key lesson that algorithmic efficiency and scalability is vital for widespread adoption.

Kernel Based Quantile Regression: Consistency and Robustness

Andreas Christmann (Free University of Brussels)
(Based on joint work with Ingo Steinwart, Los Alamos National Laboratory)

Empirical risk minimization plays an important role in statistical machine learning. Support vector machines (SVMs) proposed by V. Vapnik and general SVMs based on arbitrary convex loss functions are leading examples.
Takeuchi et al. (2006) proposed nonparametric quantile regression based on kernels. This approach can be seen as a generalization of classical quantile regression proposed by Koenker and Bassett (1978) and described in detail by Koenker (2005).
In this talk results about learnability in the sense of L−risk consistency and robustness properties of such kernel based quantile regression estimators will be given.

References:

Christmann & Steinwart (2006). Consistency of kernel based quantile regression. Submitted.
Christmann & Steinwart (2005). Consistency and robustness of kernel based regression. Submitted.
Koenker (2005). Quantile regression. Cambridge University Press.
Koenker & Bassett (1978). Regression quantiles. Econometrica, 46, 35-50.
Schoelkopf & Smola (2002). Learning with Kernels. MIT Press.
Takeuchi, Le, Sears & Smola (2006). Nonparametric quantile estimation. JMLR, 7, 1231-1264.
Vapnik (1998). Statistical Learning Theory. Wiley, New York.

Engineering Kernel Machines

Johan Suykens (K.U. Leuven, ESAT-SCD)

Support vector machines and kernel based learning is a highly multi-disciplinary research area at the intersection of different fields as machine learning and neural networks, mathematics and statistics, pattern recognition and signal processing, systems and control, optimization and others. In this respect, an important challenge is the integrative understanding and systematic design of kernel machines. In this talk we explain how least squares support vector machines may facilitate this process as 'core problem' formulations and elementary building blocks for the design of kernel machines. It enables to:

- extend support vector machine methodologies to a wide range of problems in supervised and unsupervised learning (regression, classification, principal component analysis, canonical correlation analysis, spectral clustering) and in dynamical systems (identification of different model structures, recurrent networks, optimal control) and others.
- formulate problems in terms of constrained optimization with explicit use of regularization leading to a good generalization performance and to numerically reliable schemes.
- achieve primal and dual model representations, relevant for out-of-sample extensions and solving large scale problems
- consider weighted versions towards statistical robustness and handling general loss functions
- plug-in different loss functions and positive definite kernels
- incorporate prior knowledge through additional constraints and conceive hierarchical modelling schemes using convex optimization techniques.

We illustrate the principles and methods on applications in nonlinear system identification and bioinformatics.

Convex Optimization for the Design of Learning Machines

Kristiaan Pelckmans (K.U. Leuven, ESAT-SCD)

This talk surveys convex optimization techniques for the design of task specific learning machines. Recently, techniques of Convex Optimization (CO) play a prominent role in learning approaches, as pioneered by the work on Support Vector Machines (SVMs) and other regularization based learning schemes. Duality theory has played an important role in the development of so-called kernel machines, while the fact of uniqueness of the optimal solution has permitted theoretical as well as practical breakthroughs. A third main advantage of using CO tools in research on learning problems is that the conceptual level of the design of a learning scheme becomes nicely separated from the actual algorithm implementing this scheme (the CO solver).

From a conceptual point of view, statistical learning theory established a theoretical sound foundation for prediction and generalization. This lead to necessary and sufficient requirements for guaranteeing a machine to learn effectively from observations. This talk examplifies how to use those generic principles for designing a learning machine for the task at hand, while fitting nicely into a context of convex optimization. Specifically, we will review a.o. results on primal-dual kernel machines for additive models, kernel machines in the context of missing and censored observations, and machines incorporating structural constraints. The classical supervised case, extensions towards the unsupervised case as well as formulations in the context of nonlinear dynamical systems will be discussed. We proceed by reviewing the role convex optimization can play in problems of model selection tasks. Finally, the talk will outline a generic setting for designing learning machines in the context of graph labeling tasks.

Some thoughts on kernels

Bernhard Schoelkopf (Max Planck Institute for Biological Cybernetics)

My plan is to present my thoughts on what made kernel machines popular and what may or may not keep them going. I will also discuss applications in different domains, including computer graphics.

Structure Prediction
Ioannis Tsochantaridis (Google)

Learning general functional dependencies over arbitrary inputs and outputs is one of the key challenges in machine learning. Recent progress in kernel methods has focused on designing flexible and powerful input representations, greatly increasing the range of applicability of learning methods to input representations based on strings, graphs, automata, etc. However, the output variables in supervised learning have typically been restricted to be either binary or real-valued. There are many real-world applications, nevertheless, where the outputs to be predicted have some meaningful internal structure (e.g. sequences, trees, graphs), or where multiple correlated outputs should be jointly predicted. Such problems, commonly called structure prediction problems, have recently drawn particular attention. Several approaches have been proposed and applied to problems in areas such asinformation retrieval, natural language processing, bioinformatics etc. In this talk I will give an overview of structured prediction, focusing on a specific generalization of support vector machines, and discuss challenges posed by practical applications.

How to teach the Support Vector Machine to learn vectors and structured outputs at one-class complexity
Sandor Szedmak (Southampton University, ECS, ISIS Group, and, University of Helsinki, Department of Computer Science)

The Support Vector Machine (SVM) has proved it is a very useful device of machine learning, but it was restricted to directly solve binary classification problems only. There is a strong demand to extend the underlying ideas towards multiclass classification, multivariate regression problems and learning when the outputs have complex structures, e.g. graphs. The known approaches are tackling with the exploding computational complexity and the range of potential applications is generally very limited. We show there is a straightforward algebraic generalization of the SVM which can handle arbitrary vector outputs and preserves the same computational complexity of its binary ancestor. The multiclass classification, regression and structural learning problems can then be solved via an embedding into a properly chosen vector space.

Convergence of AdaBoost and related algorithms to maximum margin solutions

Ingrid Daubechies (Princeton University)
(Based on work joint with Cynthia Rudin and R. Shapire)

The algorithm AdaBoost, invented by Freund and Shapire, has shown to be useful for a variety of applications; it constructs, out of an ensemble of weak classifiers, an appropriate combination that functions as a strong classifier. It often generalizes surprisingly well. Until a few years ago, it wasn't known whether AdaBoost converges to the solution with maximum margin when this is is strictly positive, i.e. for those classification problems where an effective strong classifier can be constructed from the weak classifiers. In her Ph. D. thesis, Rudin showed (by counterexample) that AdaBoost does not necessarily converge to the maximum margin solution; together with Shapire and the speaker (her co-advisers), she also introduced "smoothened" versions of AdaBoost, and studied their mathematical properties. In particular, they do converge to the maximum margin solution, and both their asymptotic rate of convergence and an upper bound on the number of iterations necessary to achieve a pre-assigned maximum error can be given explicitly.

Regularization path algorithms

Alain Rakotomamonjy (INSA de Rouen)

In most machine learning algorithms, a trade-off between an empirical error and a regularizer (measuring how "simple" the decision function is) has to be determined in order to obtain a predictor with good generalization performance. The objective is then to minimize both the error and the regularizer, and the interest is to find the set of classifiers denoted as the accuracy-regularization frontier that cannot reduce both terms. The typical approach for assessing the trade-off parameter is to learn several functions belonging to the "frontier" and to evaluate them with validation data or cross-validation. Recently, approaches for efficiently computing the full frontier have been proposed. The aim of this talk is to review these recent contributions.

We first present different characterizations of this frontier and then review recent methods to obtain the so-called regularization path, which is the set of classifiers belonging to the frontier achieved by an algorithm when varying the trade-off parameter. We discuss the computation of the regularization path for several specific methods for which the regularization path can be computed very efficiently, such as SVM and Least Angle Regression.

The Analysis of Patterns

Nello Cristianini (University of Bristol, Dept. Engineering Mathematics and Dept. of Computer Science)

Modern society relies on our capability to automatically detect very subtle patterns in very large and complex sets of data. This affects the way we do science, business and technology. This also poses important ethical questions. It is interesting to look at how science handles the problem of pattern analysis, and how many different scientific communities have ended up converging onto a single stable set of principles and methods for computational approaches to the same problem. Now computers can be trusted to do this job for us. How do they do it? We give a look at various approaches, as we focus on the field of kernel methods as a leading example.

Posters

On the Stability and Bias-Variance Analysis of Kernel Matrix Learning

V. Vijaya Saradhi, Harish Karnick
Dept. of Computer Science and Engg., Indian Institute of Technology Kanpur, India.

Stability and bias-variance analysis are two powerful tools for understanding of learning algorithms better. We use these tools for the analysis of learning the kernel matrix (LKM) algorithm. Motivation for analyzing the LKM comes from two directions (i) LKM works in the transductive setting where both training and test data points are to be given apriori. Hence, it is interesting to understand the sensitivity of LKM under small variations in the data set and (ii) LKMs were argued to overfit the given data set. In particular we are interested in answering the following questions (a) *Is LKM a stable algorithm?* (b) *For a give data set, what is the nature of learning for LKMs?* We use the stability and bias-variance analysis tools to answer these two questions and compare LKMs with SVMs. We consider three cases for the optimal kernel matrix: composing of (1) only Gaussian functions (2) only polynomial functions and (3) mixture of Gaussian and polynomial functions. We compare these three cases on 5 real world and artificial data sets. Our experimental results show that LKMs overfits the given data set. From the stability analysis, we found LKMs to be *unstable* algorithms even when the variance is close to 0 and vice versa. This is clearly a contradiction from the stability analysis. The primary reason for this contradiction is attributed to overfitting problem. On the bias-variance analysis front, we compare the nature of learning for LKMs under different optimal kernel cases.

A Kernel-based Approach to Retrieve Images from Text Queries

David Grangier, Samy Bengio
IDIAP Research Institute

The Passive-Aggressive Model for Image Retrieval (PAMIR) model is a discriminative model for the retrieval of pictures from text queries. This approach contrasts with alternative solutions to this multi-modal problem which typically consist of generative auto-annotation models. Compared to these approaches, PAMIR has shown to be more effective over the benchmark Corel corpus (Grangier and Bengio, ECML'06). In this work, we propose to extend the core parameterization of PAMIR, i.e. a linear mapping from the picture space to the textual space, to model non-linear mappings relying on kernels. For that purpose, we first show how the kernel trick can be applied in the model parameterization and in its learning procedure. Then, different experiments performed relying on several kernels recently introduced to compare pictures represented as sets of local descriptors. These kernels notably include the expected likelihood kernel over GMM distributions or the match (pseudo-)kernel.

Semi-supervised learning by search of optimal target vector

Daniele Marinazzo, Leonardo Angelini, Mario Pellicoro, Sebastiano Stramaglia
University of Bari and INFN

We introduce a semi-supervised learning estimator which tends to the first kernel principal component as the number of labelled points vanishes. Our approach is based on the notion of optimal target vector, which is defined as follows. Given an input data-set of $\mathbf{x}$ values, the optimal target vector $\mathbf{y}$ is such that treating it as the target and using kernel ridge regression to model the dependency of y on $\mathbf{x}$, the training error achieves its minimum value. For an unlabeled data set, the first kernel principal component is the optimal vector. In the case one is given a partially labeled data set, still one may look for the optimal target vector minimizing the training error. We use this new estimator in two directions. As a substitute of kernel principal component analysis, in the case one has some labeled data, to produce dimensionality reduction. Second, to develop a semi-supervised regression and classification algorithm for transductive inference. We show application of the proposed method in both directions.

The Volume Under the ROC Surface as Evaluation Criterion for Ordinal Regression Models

Willem Waegeman, Bernard De Baets, Luc Boullart
Ghent University

Nowadays the area under the receiver operating characteristics (ROC) curve, which corresponds to the Wilcoxon-Mann-Whitney test statistic, is increasingly used as a performance measure for binary classification systems. In this article we present a natural generalization of this concept for more than two ordered categories, a setting which is known as ordinal regression. Our extension of the Wilcoxon-Mann-Whitney statistic now corresponds to the volume of an r-dimensional surface for r ordered categories and does not suffer from shortcomings characterizing other performance measures for ordinal regression systems when the class or cost distributions are skewed. Furthermore, we present some inintial ideas on how this measure can be optimized by a kernel based quadratic program leading to models that are more appropriate for skewed cost or class distributions.

Evaluation of Performance Measures for SVM Hyperparameter Selection

Koen Smets[(1)], Brigitte Verdonk[(1)], Elsa M. Jordaan[(2)]
(1) University of Antwerp, (2) Dow Chemical Company

To obtain accurate modeling results, it is of primal importance to find optimal values for the hyperparameters in the SVM model. In general, we search for those parameters that minimize an estimate of the generalization error.

In this study, we empirically investigate different estimates found in literature: the computationally intensive, but almost unbiased leave-one-out error, its upperbounds – radius/margin and span bound – and last but not least Vapnik's measure, which uses an estimate of the VC-dimension. For each of the estimates we focus on accuracy, complexity and the presence of local minima. The latter significantly influences the applicability of gradient-based search techniques to determine the optimal parameters.

For classification tasks, the radius/margin bound in combination with a gradient-based search technique seems to provide us with an excellent tool to select the model parameters in a fast and accurate way. It opens doors that would remain closed using traditional brute force search strategies. The consideration of more complex kernels with a higher number of parameters, for example weighted kernels or mixing of multiple kernels, becomes computationally attractive.

Whether the same observations continue to hold for regression problems is the topic of ongoing research.

Fast Variable Selection for Microarray Data using Least Squares Support Vector Machines

Fabian Ojeda, Johan A.K. Suykens and Bart De Moor
Katholieke Universiteit Leuven. ESAT-SCD Division

Least squares support vector machines (LS-SVM) classifiers are a class of simple, yet powerful, kernel methods whose solution follows from a set of linear equations. Here, forward and backward algorithms for fast and efficient sequential variable selection based on this technique are proposed. By exploiting the structure of such solution a closed form expression for the leave-one-out (LOO) estimator, useful for the selection of relevant variables, is obtained. For inclusion or removal of a new variable, rank-one adjustments in the kernel matrix allow for updating, rather than recomputing, the LS-SVM solution. The risk of highly overfitted models, given the large number of variables relative to the number of samples in microarray data, is reduced when an external cross-validation procedure is employed. Simulations clearly show lower computational complexity along with better stability on the generalization performance when compared to other related algorithms.

An Out-of-Sample Extension to Spectral Clustering based on Weighted Kernel PCA and its Application to Image Segmentation

Carlos Alzate, Johan A. K. Suykens
ESAT-SCD-SISTA, K.U.Leuven

Spectral clustering techniques have emerged as promising unsupervised learning methods to group data points that are similar. These methods have been successfully applied to machine learning, data analysis, image processing, pattern recognition and VLSI design.

Starting from graph theory, spectral clustering methods are formulated as optimization problems in which the objective is to minimize the cost of a partitioning (also called cut) of a graph representing the data. In general, these optimization problems are NP-hard due to a discrete constraint on the clustering indicator matrix. Approximate solutions can be obtained using an eigenvalue decomposition of the Laplacian matrix. The eigenvectors become the relaxed solutions. In general, spectral clustering techniques are formulated only for the training set and extensions to out-of-sample data points should rely on approximation algorithms such as the Nystrom method.

Image segmentation is a fundamental problem in image processing and computer vision. It consists of partitioning an image into several constituent sets based on some criterion. Typically, it is used to distinguish objects from the background. In graph-based image segmentation, every pixel becomes a vertex in the graph and pairwise similarities are computed between each pair of pixels. Therefore, the size of the Laplacian matrix is rather large even for small images.

Kernel Principal Component Analysis (Kernel PCA) is an unsupervised learning technique used for nonlinear feature extraction, dimensionality reduction and denoising. The objective is to find projected variables with maximal variance in a kernel induced feature space. The solutions are given by the eigenvectors of the centered kernel matrix of a dataset and the corresponding eigenvalues indicate the amount of variance captured by the projection. A weighted formulation to kernel PCA based on the Least Squares Support Vector Machines (LS-SVM) framework has been developed. This weighted scheme was used as a fast and efficient technique to impose robustness and sparseness in kernel PCA.

In this work, we consider a different kind of weighting which leads to a new formulation to spectral clustering. This formulation allows us to extend the trained model to out-of-sample data points by means of the projections onto the eigenvectors and without relying on approximations of the underlying eigenfunction such as Nystrom. The proposed out-of-sample extension is based on the primal-dual formulation insights given by the LS-SVM framework. The clustering model is first trained on a subsampled version of the image and then the cluster indicators of the remaining pixels are inferred by means of the out-of-sample extension. In this way, the computational complexity is reduced considerably.

Probabilistic Inference, Covariate Shift and Reduction by Touchstone Class

Tsuyoshi Okita, Bernard Manderick
Vrije Universiteit Brussel

Opposed to the Bayesian approach (*distribution*-based approach), traditional statistical learning tends to avoid the calculation of the underlying unknown but fixed distribution, but evalutes the criteria directly by training examples (*point*-based approach). Typical examples are SVMs and other kernel methods. And indeed this strategy led to the victory of SVMs over other methods in classification problems. However recently several persons come to believe that in some categories of problems it might be advantageous to know this unknown distribution. Touchstone class (Shawe-Taylor, De Bie, and Dolia, 2006) is invented for such general purposes where this class aims to provide an opportunity to access the underlying distribution during its calculation for evaluating the criteria in a frequentist manner. This poster focuses on several new perspectives that this class may open for statistical learning: probabilistic inference, covariate shift (where the input and the output distribution are different), and reduction (where we consider any types of output distribution). In terms of kernels in this context we highlight a kernel induced by a nuclear operator, where this nuclear operator resides in the bottom of the stratified structure of operators in Hilbert space (compact operator $\subseteq$ HS operator $\subseteq$ nuclear operator).

Robustness of reweighted kernel based regression

Michiel Debruyne[(1)], Andreas Christmann[(2)], Mia Hubert[(1)], Johan Suykens[(3)]
(1) K.U.Leuven, department of mathematics - university center for statistics, (2) Vrije Universiteit Brussel, department of mathematics, (3) K.U.Leuven, ESAT-SCD/SISTA

Recently it was shown that Kernel Based Regression (KBR) with a least squares loss function may have some undesirable properties from a robustness point of view: even very small amounts of outliers can dramatically affect the estimates. KBR with other loss functions is much more robust, but often gives rise to more complicated computations. In classical statistics robustness is often improved by reweighting the original estimate. In this poster we provide a theoretical framework for reweighted KBR and analyze its robustness. Some important differences are found with respect to linear regression, indicating that LS-KBR with a bounded kernel is much more suited for reweighting. Our results give practical guidelines for a good choice of weights. In particular a logistic weight function seems an appropriate choice. Robustness is analyzed by means of the influence function. Some interesting links with stability and the leave-one-out-error are explored as well, indicating that reweighting also improves the stability of KBR, especially at heavy-tailed distributions.

Diagnosis of complex system by combined use of RKCCA and Graphical Model

Etienne Côme, Latifa Oukhellou and Patrice Aknin
INRETS

The aim of our study is to perform a diagnosis of a complex system made up of several subsystems that are spatially dependent along a direction: a defective subsystem modifies the inspection data related to downstream subsystems. In addition, the number of these subsystems is not always the same. This leads us to develop a generative model which performs the diagnosis of all subsystems by taking into account these particularities: spatial dependency, unidirectional causality and variable space dimension.

During a preliminary study, we have obtained interesting results with 'Linear / Gaussian' relationship between the nodes of a graphical model which describes the generative model. Afterwards, a non-linear behaviour of the conditional relationship between nodes has been noticed. The classical approaches to solve such problem are often based on sampling technique and are time consuming. So, we propose the use of 'Regularized Kernel Canonical Correlation Analysis (RKCCA)' as a pre-processing step to learn the non-linear relationship between nodes of the graphical model. RKCCA is used to find a common space where the relation between the nodes became linear, so that we may learn it with a classical 'linear / Gaussian' model.

The approach is illustrated on a railway device diagnosis application. The performances highlight the advantage of the combined use of these two methods (RKCCA + Graphical model) in order to take into account all the available information. We will also discuss some practical aspects of such method as the design of a suitable kernel, and the tuning of the regularization parameter.

Improved core promoter prediction using ensembles of support vector machines

Thomas Abeel, Yvan Saeys and Yves Van de Peer
VIB - Ghent University, Department of Plant Systems Biology, Bioinformatics & Evolutionary Genomics, Technologiepark 927, B-9052 Gent, Belgium

Computer-aided gene prediction is one of the hot topics in genome analysis because it allows for computational annotation of genomes. The region before a gene is called the promoter. This promoter and in particular the core promoter is responsible for the activation of a gene. The identification of gene promoters and their regulatory elements is one of the biggest challenges in bioinformatics. The core promoter is the region close around the transcription start site (TSS) (we took a region of -200, +50 around the TSS). Machine learning techniques are often used to detect putative TSSs, although several problems exist. First, the datasets involved are rather large, ranging from several thousands of training instances for the positive data to several hundreds of thousands of negative samples. The second problem is the class imbalance between positive and negative samples. In the human genome only 3% of the sequence codes for genes and an even smaller part is located within the core promoter. This large imbalance is very difficult to model in e.g. support vector machines (SVM) as there is so little positive information.

Here, we explore a technique to reduce the training time for support vector machines, while increasing their prediction performance. The used technique is an ensemble of support vector machines, each trained on a different part of the training set. Every one of these support vector machines is then validated on a separate validation dataset which is different from the training data. This step is needed to determine the optimal number of support vector machines that must have a positive output to classify an instance as positive. While the training time of one SVM using a RBF kernel with increasingly large datasets does not increase linearly, the time needed for training of separate SVMs on random sub samples of the large dataset increases linearly when we go for single coverage of the original dataset. We selected more subsets than was strictly necessary to stand a better chance to cover the original dataset completely.

We compared the performance using different evaluation measures for both the ensemble of support vector machines and a single support vector machine. This technique was applied to a core promoter classification task. The goal of this task was to distinguish core promoters from gene and intergenic sequences. We performed our analysis on four different species: rice, arabidopsis, mouse and human. The datasets can be considered to be large, containing 2500-7000 positive training examples and about ten times more negative examples, each sample consisting of 250 features. Training on a dataset of 4500 positive and 13500 negative samples on a single SVM took nearly 4 days and gave a recall of 77%, a precision of 88% and an F-measure value of 0.82. The validation was done on the training data using a 10-fold cross validation.

Our approach with 26 SVMs each trained on 3200 samples (800 positive and 2400 negative) with a validation of 8000 sequences (80 positive, 7920 negative) performed better. The validation used is much stricter as it is much more difficult for the SVM to predict true positives. The training and validation only took a couple of hours instead of 4 days. This method obtained a recall of 94%, a precision of 92% and an F-measure value of 0.93.

This analysis clearly shows that the use of ensembles of support vector machines is superior to the use of one single SVM, both in time as in classification performance. This approach can also be very valuable for similar problems with skewed classes, like splice-site prediction.

Online Independent Support Vector Machines

Francesco Orabona, Claudio Castellini, Giulio Sandini
DIST, University of Genoa

One of the most interesting characteristics of Support Vector Machines (SVMs) is that the solution achieved during training is sparse, meaning that a few samples are usually considered ‘important’ by the algorithm (the so-called support vectors) and account for most of the complexity of the classification task. However, it is also known that often the number of support vectors can be reduced by simplifying the obtained solution, without losing accuracy. In fact, it is crucial to keep the number of support vectors as small as possible, to reduce the testing time. Low-rank approximations for kernel methods have been proposed to address this problem, but they need the knowledge of the entire training set and are not capable of on-line training.

We present a new algorithm for online SVM training with a reduced number of support vectors, exploiting the linear indipendence of the support vectors in the feature space. The algorithm breaks the linear dependency between the number of support vectors and the number of training points. Sparseness can be further increased by weakening the concept of linear indipendence, trading accuracy for sparseness.

Experiments reveal that our machines achieve a dramatic reduction in the number of support vectors without losing accuracy, especially when finite-dimensional kernels are used. Moreover, even with infinite-dimensional kernels, the dependence of the test error from the weakened linear independence is definitely bearable, the test error getting ~1% worse when 10 times fewer support vectors are retained.

List of speakers and participants

Abeel	Thomas	VIB - Ghent University
Abhilash	Miranda	Université Libre de Bruxelles, Machine Learning Group
Alzate	Carlos	ESAT-SISTA K.U.Leuven
Ancona	Nicola	Institute of Intelligent Systems for Automation
Blockeel	Hendrik	K.U.Leuven, Computer Science
Bonilla	Edwin	University of Edinburgh
Brew	Anthony	University College Dublin, Machine Learning Group
Buelens	Bart	University of Tasmania
Callut	Jerome	Universite catholique de Louvain
Camargo	Anyela	University of Ulster
Cara	Philippe	Vrije Universiteit Brussel
Chen	Yifei	VUB/COMO
Christmann	Andreas	Vrije Universiteit Brussel
Claeys	Petra	Univ. Gent
Côme	Etienne	INRETS (France)
Cristianini	Nello	University of Bristol
Crucianu	Michel	Conservatoire National des Arts et Métiers
Cunningham	Pádraig	University College Dublin
Dablemont	Simon	Université catholique de Louvain
Daelemans	Walter	University of Antwerp, CNTS
Daemen	Anneleen	K.U.Leuven, ESAT/SCD (SISTA)
d'Alché	Florence	Université d'Evry
Damien	François	Université catholique de Louvain
Daubechies	Ingrid	Princeton University
De Bie	Tijl	Katholieke Universiteit Leuven (KUL)
De Bruyne	Steven	Vrije Universiteit Brussel/MOSI
De Moor	Bart	K.U.Leuven ESAT/SCD (SISTA)
Debruyne	Michiel	KULeuven
Delannay	Nicolas	Université catholique de Louvain
Dhollander	Thomas	K.U.Leuven
DiFranco	Matthew	University College Dublin
Dineen	David	School of Computer Science and Informatics, UCD
Driessens	Kurt	K.U.Leuven
Dupont	Pierre	Université catholique de Louvain
Espinoza	Marcelo	Katholieke Universiteit Leuven
Farida	Zehraoui	IBISC
Fouss	François	Université catholique de Louvain
Fromont	Elisa	K.U.Leuven, Computer Science
Gallo	Arianna	univ of torino / univ of Bristol
Garcia-Gomez	Juan M	Universidad Politecnica de Valencia

Gasso	Gilles	LITIS, INSA de Rouen
Geurts	Pierre	University of Liège
Glasmachers	Tobias	Ruhr-Universität Bochum, Neuroinformatik
Gómez Verdejo	Vanessa	Université catholique de Louvain
Gordillo	José	Vrije Universiteit Brussel/MOSI
Grangier	David	IDIAP Research Institute
Greene	Derek	University College Dublin
Grobelnik	Marko	Jozef Stefan Institute (Slovenia)
Guiza	Fabian	K.U.Leuven, Computer Science
Helleputte	Thibault	Université catholique de Louvain
Hillewaere	Ruben	Vrije Universiteit Brussel
Hubert	Mia	Katholieke Universiteit Leuven
Igel	Christian	Ruhr-Universität Bochum, Neuroinformatik
Iria	Jose	University of Sheffield, Computer Science
Jordaan	Elsa	Dow Benelux N.V.
Karsmakers	Peter	K.U.Leuven
Lanzarotti	Raffaella	Dipartimento di Scienze dell'Informazione
Lipori	Giuseppe	Università degli Studi di Milano
Liu	Feng	VUB/COMO
Luh	Yen	Université Catholique de Louvain
Luts	Jan	Katholieke Universiteit Leuven
Maglietta	Rosalia	CNR Issia (Italy)
Manderick	Bernard	Vrije Universiteit Brussel (VUB)
Marinazzo	Daniele	Università and INFN Bari, Fisica
Markovsky	Ivan	K.U.Leuven ESAT/SISTA
Marttinen	Pekka	University of Helsinki
Mladenic	Dunja	Jozef Stefan Institute (Slovenia)
Moreau	Yves	K.U.Leuven, ESAT/SCD (SISTA)
Nessler	Bernhard	Technische Universität Graz
Nicola	Cancedda	Xerox Research Centre Europe
Nowé	Ann	Vrije Universiteit Brussel (VUB)
Ojeda	Fabian	Katholieke Universiteit Leuven
Okita	Tsuyoshi	Vrije Universiteit Brussel
Onclinx	Victor	Université Catholique de Louvain-la-Neuve
Orabona	Francesco	University of Genoa
Pelckmans	Kristiaan	Katholieke Universiteit Leuven
Pellicoro	Mario	Università and INFN Bari, Fisica
Plastria	Frank	Vrije Universiteit Brussel
Ponsen	Marc	Universiteit Maastricht
Rademaker	Michael	Ugent
Rakotomamonjy	Alain	INSA de Rouen
Ramon	Jan	K.U.Leuven, Computer Science
Ricci	Elisa	univ of perugia / univ of bristol
Saeys	Ivan	Ghent University, Bioinformatics group
Sagaert	Steven	VUB/COMO

Schietgat	Leander	Katholieke Universiteit Leuven
Schölkopf	Bernhard	Max Planck Institute for Biological Cybernetics
Shawe-Taylor	John	University College London
Slabbinck	Bram	Ghent University
Smeets	Hugues	VUB
Smets	Koen	University of Antwerp
Spickenreuther	Thorsten	University of Bayreuth
Stobbe	Miranda	AMC Medical Research
Stramaglia	Sebastiano	Università and INFN Bari, Fisica
Struyf	Jan	K.U.Leuven, Computer Science
Suykens	Johan	Katholieke Universiteit Leuven
Szedmak	Sandor	University of Southampton
Tomioka	Ryota	Fraunhofer FIRST.IDA
Trafalis	Theodore	The University of Oklahoma
Tsochantaridis	Ioannis	Google
Tuerlinckx	Francis	KULeuven, Quantitative Psychology
Tuyls	Karl	Maastricht University
Uwents	Werner	K.U.Leuven, Computer Science
Van den Bulcke	Tim	KULeuven, ESAT-SCD
van der Maaten	Laurens	Universiteit Maastricht
Van Messem	Arnout	Vrije Universiteit Brussel
van Oeffelen	Liesbeth	K.U.Leuven, ESAT/SCD (SISTA)
Vanderlooy	Stijn	MICC-IKAT, Maastricht University
Vanschoren	Joaquin	K.U.Leuven, Computer Science
Verbeeck	Katja	VUB
Verena	Heidrich-Meisner	Ruhr-Universität Bochum, Neuroinformatik
Verleysen	Michel	Université catholique de Louvain, Machine Learning Group
Vert	Jean-Philippe	Ecole des Mines de Paris
Vijaya Saradhi	Vedula	Indian Institute of Technology, Kanpur
Villalba	Santiago David	Machine Learning Group, UCD
Waegeman	Willem	Ghent University