Machine Learning, 46, 5–9, 2002

Editorial: Kernel Methods: Current Research and Future Directions

The introduction of Support Vector Machines (SVMs) in COLT 1992 (Boser, Guyon, & Vapnik, 1992) has been an important innovation in the field of Machine Learning. In addition to their accuracy, a key characteristic of SVMs is their mathematical tractability and geometric interpretation. This has facilitated a rapid growth of interest in SVMs, resulting in the underlying concepts and techniques being generalized and formalized, and shedding light on new connections with other approaches.

SVMs have been successfully extended from basic classification tasks to handle regression, operator inversion, density estimation, novelty detection, and to include other desirable traits, such as invariance under symmetries and robustness in the presence of noise. In fact, the entire research direction in learning theory concerned with large margin classification can be traced back to SVMs (Shawe-Taylor et al., 1998; Vapnik, 1998). Support Vector Machines (along with Adaboost) have, in fact, been one of the few algorithmic approaches originally motivated by learning theory and subsequently introduced into the standard toolbox of practitioners.

Since the inception of this subject, a number of systems have been recognized (or adapted) to implement a similar learning bias (Schapire et al., 1998; Bennett et al., 2000). Furthermore, linear algorithms can be converted into nonlinear algorithms through the use of kernel substitution (implicitly mapping data into a high-dimensional feature space) and this idea has been successfully applied to other algorithms, such as PCA, clustering, and Bayesian classifiers. At the same time it has been recognized that regression and classification systems based on Gaussian Processes (Williams, 1998) are in fact kernel-based learning systems and are therefore closely connected to SVMs.

The result is that hundreds of articles have been written, extending in one way or another the ideas of that first paper, and creating a research field that has come to be known as Kernel-Based Learning Methods, or Kernel Methods (KMs) for short. A website jointly managed by an editorial board acts as a coordinating node for this new research community (**www.kernel-machines.org**), hosting papers, software, a list of researchers, and links.

Books have begun to appear in the last few years (Cristianini & Shawe-Taylor, 2000; Schölkopf & Smola, to be published) providing a unified foundation for this heterogeneous field. Since 1997 an annual workshop has been held regularly at the Neural Information Processing Systems (NIPS) conference, and the papers of the 1997 and 1998 workshops have been collected in edited books (Schölkopf, Burges, & Smola, 1999; Smola et al., 2000).

Continuing this tradition, this special issue is mostly (but not exclusively) based on the papers presented in the 1999 NIPS workshop on kernel methods.

After attracting considerable interest from the more theoretically-inclined research community, kernel methods are now also part of the toolbox of many practitioners. There are several reasons for the success of this new class of learning methods. The main one is certainly a series of remarkably successful applications in fields as diverse as text categorization, bioinformatics and machine vision. SVMs do work very well in practice, as witnessed, for example, by one of the articles in this issue, in which the best current performance on the MNIST digit recognition benchmark is demonstrated.

Another reason for their success stems from their appealing properties. One is their modularity: any kernel-based learning algorithm can work with any kernel function and vice versa. In this way one can separate the two tasks of feature selection (part of the kernel design effort) and learning (part of the learning machine design). The kernel functions themselves can be derived in a modular way, combining simple kernels to obtain complex ones.

The hypothesis constructed by most KMs depends only on a subset of the training data points, called "support vectors", which can also be viewed as the most informative training patterns. Removal of the non-support vectors and re-training results in the same solution. In many cases the support vectors form a small subset of the training data, resulting in a *sparse* solution, although methods to significantly reduce the number of effective support vectors ("reduced set methods"), and thus greatly increase speed in test phase, have also been invented (Burges, 1996). It turns out that the property of sparseness has consequences both for learning theory (compression bounds) and for implementation.

Additional advantages of this approach can be appreciated in comparion to neural networks. For SVMs there are only a small number of tunable parameters, and training amounts to solving a convex quadratic programming problem hence giving solutions that are global, and usually unique (Burges & Crisp, 2000). The absence of local minima is a significant benefit during the learning process, with the learning parameters converging monotonically towards the solution. In addition the architecture of the learning machine is given by the algorithm. For all these reasons, kernel methods have become an increasingly popular alternative to neural network approaches.

Most of the papers in this Special Issue were initially presented as talks at the Neural Information Processing (NIPS) Workshop on Support Vector Machines and other Kernel Based learning methods held in Breckenridge, Colorado in December, 1999. Broadly, the papers can be divided into theoretical analysis, the development of new algorithms, and new applications. Theoretical innovations are presented using ideas from statistical learning theory, Bayesian analysis and statistical mechanics. New algorithms are presented for improving generalization performance in addition to efficient implementations that can scale up to hundreds of thousands of datapoints. Finally, new application studies are presented for text categorization, digit recognition and bioinformatics.

In the first paper of the volume Chris Williams (*On a connection between kernel PCA and metric multidimensional scaling*) shows that, for some kernel functions, Kernel PCA algorithm can be interpreted as a form of multi-dimensional scaling. This leads to a new algorithm for multidimensional scaling based on eigencomputations. Peter Sollich (*Bayesian Methods for Support Vector Machines: Evidence and Predictive class probabilities*) provides

a framework for interpreting SVMs as maximum a posteriori (MAP) solutions to inference problems with Gaussian process priors. This not only gives an alternative interpretation of the SVM algorithm but also provides new approaches to model selection and tuning. Sebastian Risau-Gusman and Mirta Gordon (*Hierarchical Learning in polynomial Support Vector Machines*) use methods from Statistical Mechanics to investigate the generalization properties of a class of polynomial SVMs. Whereas arguments from statistical learning theory give loose upper bounds on generalization, this approach can give new insights for the averaged generalization error of SVMs.

The paper *A Probabilistic Framework for SVM Regression and Error Bar Estimation* by Junbin Gao, Steve Gunn and Chris Harris derives an expression for the evidence and an error bar approximation formula for regression problems. Tong Zhang in *On the Dual Formulation of Regularised Linear Systems with Convex Risks"* presents a general approach to linear prediction algorithms with a number of known schemes given as special cases. The paper *Choosing Multiple Parameters for Support Vector Machines*, by Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet and Shayan Mukherjee addresses the problem of automatically tuning parameters for pattern recognition SVMs. This is done by minimizing some leave-one-out estimates of the generalization error by gradient descent.

Dennis DeCoste and Bernhard Schoelkopf (*Training Invariant Support Vector Machines*) review several known methods for incorporating prior knowledge about invariances into SVMs. They report important new results with a new performance record on the standard MNIST benchmarking dataset for handwritten digits recognition (the lowest reported test error) with SVM training times significantly faster than previous SVM methods. Yi Lin, Yoonkyung Lee and Grace Wahba (*Support Vector Machines for Classification in Nonstandard Situations*) consider the problem of learning when the cost of misclassification is different in the two classes. Theodore Trafalis and Alexander Malyscheff (*An Analytic Center Machine*) propose a new algorithm with reported superior generalization performance over SVMs. Ayhan Demiriz, Kristin Bennett and John Shawe-Taylor (*Linear Programming Boosting via Column Generation*) present a boosting technique (LPBoost) which can be applied to any boosting task formulated as a linear programming problem. In particular they examine its use with a 1-norm soft margin cost function which can be used to train a SVM. This approach is attractive theoretically and can be readily implemented using fast column generation techniques, for example. Olvi Mangasarian and David Musicant (*Large Scale Kernel Regression via Linear Programming*) describe a new linear programming formulation of SVMs, which presents several advantages both from the point of view of scaling and of robustness to noise.

Scalability is a central problem if kernel methods are to be used on real world problems which may contain millions of datapoints. Many different approaches have been proposed to handle large datasets and improve the speed of convergence to a solution. Gary Flake and Stephen Lawrence (*Efficient SVM Regression Training with SMO*) give a generalization of the SMO algorithm of John Platt to the problem of regression. In doing so they also modify the algorithm to increase its efficiency, improving its convergence rate by an order of magnitude. Chih-Wei Hsu and Chih-Jen Lin (*A Simple Decomposition Method for Support Vector Machines*) address the problem of working-set selection for the decomposition method, a common procedure for training SVMs on large datasets. The simple

solution they provide turns out to be a powerful one, as demonstrated by the experimental results. Pavel Laskov (*Feasible Direction Decomposition Algorithm for Training Support Vector Machines*) presents a decomposition algorithm for training SVMs based on the method of feasible directions, and discusses its relations with other related algorithms. Sathiya Keerthi and Elmer Gilbert (*Convergence of a Generalized SMO Algorithm for SVM Classifier Design*) study a class of Support Vector Algorithms that generalize the simple and efficient SMO algorithm and give proof of convergence. This class of algorithms is significantly faster than the standard SMO. Yi Li and Phil Long (*The Relaxed Online Maximum Margin Algorithm*) describe a new incremental algorithm for training linear threshold functions: ROMMA. It can be viewed as an approximation to an algorithm that repeatedly chooses the hyperplane which classifies previously seen examples correctly with the maximum margin. A mistake bound, that is the same as for the perceptron, is proven. This is the first worst-case performance guarantee of maximal margin classifiers.

The *raison d'etre* of machine learning algorithms is performance on real world problems. SVMs have been successfully applied in many fields, but they seem to have been particularly successful in application to text categorization, bioinformatics and machine vision problems where the high dimensionality of the problem often prohibits the use of alternative learning techniques. Apart from the digit recognition results of Dennis DeCoste and Bernhard Schoelkopf, two further application studies are given. Isabelle Guyon, Jason Weston, Stephen Barnhill and Vladimir Vapnik (*Gene Selection for Cancer Classification using Support Vector Machines*) describe the application of SVMs to gene expression data derived from DNA microarrays, one of the fields where the application of SVMs is most promising. In particular, they present an iterative procedure for feature selection which can be used to identify highly informative genes for cancer prediction. Edda Leopold and Jorg Kindermann (*Text Categorization with Support Vector Machines: how to Represent Text in Input Space*?) apply SVMs to the important domain of text categorization, trying different representation schemes for the text documents. In particular, they report no advantage in performing the expensive step of stemming, and a strong dependence of the performance on the choice of term-weighting schemes.

Overall, the papers in this volume represent an excellent summary of the important developments currently underway in this subject. Record performance on important datasets, and the application to very diverse domains, illustrate the power of these techniques and we expect many new successful applications in coming years.

We would like to thank Rich Caruana and Sue Becker for their help with the arrangements for the NIPS Workshop. We would also like to thank the 58 referees who played an important role in the preparation of this volume.

Nello Cristianini
Colin Campbell
Chris Burges

References

Bennett, K., Cristianini, N., Shawe-Taylor, J., & Wu, D. (2000). Enlarging the margin in perceptron decision trees. *Machine Learning*, *41*, 295–313.

Bosen, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (pp. 144–152). Pittsburgh, PA: ACM Press.

Burges, C. (1996). Simplified support vector decision rules. In L. Saitta (Ed.), *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 71–77). Bari, Italy: Morgan Kaufman.

Burges, C. & Crisp, D. (2000). Uniqueness of the svm solution. *NIPS*, 12.

Cristianini, N. & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge, UK: Cambridge University Press. www.support-vector.net.

Schapire, R., Freund, Y., Bartlett, P., & Lee, W. S. (to appear). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*. An earlier version appeared in D. H. Fisher, Jr. (Ed.), *Proceedings ICML97*, Morgan Kaufmann.

Schölkopf, B., & Burges, C. J. C., & Smola, A. J. (1999). *Advances in kernel methods—Support vector learning*. Cambridge, MA: MIT Press, 1999.

Schölkopf, B. & Smola, A. J. (to be published). *Learning with kernels*. Cambridge, MA: MIT Press.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., & Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*.

Smola, A., Bartlett, P., Schölkopf, B., & Schuurmans, C. (2000). *Advances in large margin classifiers*. Cambridge, MA: MIT Press.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Williams, C. K. I. (1998). Prediction with gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan (Ed.), *Learning and inference in graphical models*. Dordrecht: Kluwer.

Notation used

K: Mercer kernel e.g. $K(x_i,\ x_j)$
F: feature space ($\phi(x)$ is mapping function to feature space).
n: dimensionality of input space
x_i: input patterns
y_i: labels or target values.
ℓ: number of training examples
w: weight vector
b: bias in decision function
d: VC dimension
α_i: Lagrange multiplier
ξ_i: slack variables
H: Hessian for quadratic programming
L: primal lagrangian
W: dual lagrangian