Seong-Whan Lee
Alessandro Verri (Eds.)

LNCS 2388

Pattern Recognition with Support Vector Machines

First International Workshop, SVM 2002
Niagara Falls, Canada, August 2002
Proceedings

Springer

Seong-Whan Lee Alessandro Verri (Eds.)

Pattern Recognition with Support Vector Machines

First International Workshop, SVM 2002
Niagara Falls, Canada, August 10, 2002
Proceedings

Springer

Volume Editors

Seong-Whan Lee
Korea University, Department of Computer Science and Engineering
Anam-dong, Seongbuk-ku, Seoul 136-701, Korea
E-mail: swlee@image.korea.ac.kr

Alessandro Verri
Università di Genova
Dipartimento di Informatica e Scienze dell'Informazione
Via Dodecaneso 35, 16146 Genova, Italy
E-mail: verri@disi.unige.it

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Pattern recognition with support vector machines : first international workshop ; proceedings / SVM 2002, Niagara Falls, Canada, August 10, 2002. Seong-Whan Lee ; Alessandro Verri (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2388)
ISBN 3-540-44016-X

CR Subject Classification (1998): I.5, I.4, F.1.1, F.2.2, I.2.6, I.2.10, G.3

ISSN 0302-9743
ISBN 3-540-44016-X Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York,
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by DA-TeX Gerd Blumenstein
Printed on acid-free paper SPIN: 10870546 06/3142 5 4 3 2 1 0

Preface

With their introduction in 1995, Support Vector Machines (SVMs) marked the beginning of a new era in the learning from examples paradigm. Rooted in the Statistical Learning Theory developed by Vladimir Vapnik at AT&T, SVMs quickly gained attention from the pattern recognition community due to a number of theoretical and computational merits. These include, for example, the simple geometrical interpretation of the margin, uniqueness of the solution, statistical robustness of the loss function, modularity of the kernel function, and overfit control through the choice of a single regularization parameter.

Like all really good and far reaching ideas, SVMs raised a number of interesting problems for both theoreticians and practitioners. New approaches to Statistical Learning Theory are under development and new and more efficient methods for computing SVM with a large number of examples are being studied. Being interested in the development of trainable systems ourselves, we decided to organize an international workshop as a satellite event of the 16th International Conference on Pattern Recognition emphasizing the practical impact and relevance of SVMs for pattern recognition.

By March 2002, a total of 57 full papers had been submitted from 21 countries. To ensure the high quality of workshop and proceedings, the program committee selected and accepted 30 of them after a thorough review process. Of these papers 16 were presented in 4 oral sessions and 14 in a poster session. The papers span a variety of topics in pattern recognition with SVMs from computational theories to their implementations. In addition to these excellent presentations, there were two invited papers by Sayan Mukherjee, MIT and Yoshua Bengio, University of Montreal.

SVM 2002 was organized by the Center for Artificial Vision Research at Korea University and by the Department of Computer and Information Science at the University of Genova. We wish to thank all the members of the Program Committee and the additional reviewers who managed to review the papers in a very short time. We are also grateful to Sang-Woong Lee for developing and maintaining the wonderful web-based paper submission/review system. Finally we thank our sponsors, the Center for Biological and Computational Learning at MIT, the Brain Science Research Center at KAIST, the Statistical Research Center for Complex Systems at Seoul National University, and WatchVision, Inc. for their support.

We hope that all presenters and attendees had an enjoyable SVM 2002. There will have been ample time for discussion inside and outside the workshop hall and plenty of opportunity to make new acquaintances. Last but not least, we would like to express our gratitude to all the contributors, reviewers, program committee members, and sponsors, without whom the workshop would not have been possible.

May 2002

Seong-Whan Lee
Alessandro Verri

Workshop Co-chairs

S.-W. Lee Korea University, Korea
A. Verri University of Genova, Italy

Program Committee

H. Bischof	Vienna University of Technology, Austria
C.J.C. Burges	Lucent Technologies, USA
H.-R. Byun	Yonsei University, Korea
G. Cauwenberghs	Johns Hopkins University, USA
N. Cristianini	University of London, UK
X. Ding	Tsinghua University, China
R.P.W. Duin	Delft University of Technology, The Netherlands
S. R. Gunn	University of Southampton, UK
I. Guyon	ClopiNet, USA
B. Heisele	Honda R&D America, USA
S. S. Keerthi	National University of Singapore, Singapore
J. Kittler	University of Surrey, UK
A. Leonardis	National Taiwan University, Taiwan
S. Mukherjee	MIT, USA
J. Park	Korea University, Korea
P. J. Philips	NIST, USA
P. Perner	IBal Leipzig, Germany
I. Pitas	University of Thessaloniki, Greece
J. Platt	Microsoft Research, USA
H. Shimodaira	JAIST, Japan
A. J. Smola	Australian National University, Australia
H. Taira	NTT, Japan
K. Tsuda	AIST, Japan
V. Vapnik	AT&T, USA
Q. Zhao	University of Florida, USA

Organized by

Center for Artificial Vision Research, Korea University
Computer Science Department, University of Genova

Sponsored by

Brain Science Research Center, KAIST
Center for Biological and Computational Learning, MIT
Statistical Research Center for Complex Systems, Seoul National University
WatchVision, Inc.

In Cooperation with

IAPR TC-1
IAPR TC-11
IAPR TC-17

Table of Contents

Invited Papers

Computational Issues

Object Recognition

Pattern Recognition

Applications

Poster Papers

Predicting Signal Peptides with Support Vector Machines

Neelanjan Mukherjee[1,2] and Sayan Mukherjee[1,3]

[1] Center for Biological and Computational Learning
Massachusetts Institute of Technology, 45 Carleton St., Cambridge, MA 02139
nmukherj@ucsd.edu
sayan@mit.edu
[2] Uinversity of California San Diego, Dept. of Biology
[3] Center for Genome Research, Whitehead Institute
Massachusetts Institute of Technology

Abstract. We examine using a Support Vector Machine to predict secretory signal peptides. We predict signal peptides for both prokaryotic and eukaryotic signal organisms. Signalling peptides versus non-signaling peptides as well as cleavage sites were predicted from a sequence of amino acids. Two types of kernels (each corresponding to different metrics) were used: hamming distance, a distance based upon the percent accepted mutation (PAM) score trained on the same signal peptide data.

1 Introduction

For both prokaryotic and eukaryotic cells, proteins are transported from their cite of synthesis to other cites either inside or outside the cell. A basic step in the transportation process is to mark proteins for translocation across membranes: e.g. cell membrane, outer membrane, and endoplasmic recticulum (ER). The protein destination depends on the sequence of amino acids located at the n-terminus of the nascent protein chain bound to the ribosome. This sequence or targeting signal is called a signal peptide (SP).

Discriminating a signal peptide from a non-signal peptide or finding the location of the cleavage site between the two is of practical importance because of the need to find more effective vehicles for protein production in recombinant systems. It is thought that cells recognize signal peptides with almost 100% selectivity and specificity [1]. Signal peptides do have particular characteristics that are consistent for eukaryotic and prokaryotic cells. One characteristic is that signal peptides can typically be separated into three regions. Other characteristics relate to the frequency of occurrence of particular amino acids at particular locations along the sequence. However, because the signal peptides do not have unique consensus sequences these biological characterizations do not provide an accurate enough classification rule.

Pattern recognition algorithms maybe appropriate for this problem since there exists a large set of examples from which to infer a set of rules which

S.-W. Lee and A. Verri (Eds.): SVM 2002, LNCS 2388, pp. 1–7, 2002.

discriminate between two patterns, either signal peptides vs. non-signal peptides or cleavage site vs. non-cleavage site. In the past neural networks, Hidden Markov Models (HMMs), and neural networks coupled with HMMs [2,1,3] were used for the discrimination. In the paper we explore using SVMs for the discrimination. The reasons for using an SVM are as follows: for a variety of problems SVMs have performed very well [4], unlike a neural network the SVM might give some interesting biological feedback upon examining the protein sequences of the margin SVs (the examples that determine the discrimination boundary), and an HMM can be embedded in an SVM [5] avoiding ad hoc algorithms used to couple the neural networks and HMMs.

The paper is organized as follows. Section 2 gives some background about what is known about signal peptides for prokaryotes and eukaryotes and describes the data. Section 3 introduces SVMs and the two types of kernels or distance metrics used. Section 4 describes the results of our prediction algorithms and compares them to other studies.

2 Signal Peptide Properties and Datasets

In this section we summarize some characteristics of signal peptides for eukaryotic and prokaryotic cells and the datasets used. In both types of cells the signal sequence can be separated into three regions: a positively charged n-region followed by a hydrophobic h-region and a neutral but polar c-region. The $(-3, -1)$ rule states that the residues at position -3 and -1, relative to the cleavage site, must be small and neutral for cleavage to occur. We will look at both eukaryotic and prokaryotic cells.

The dataset was the same as that used by [1] which was taken from SWISS-PROT version 29 [6]. The dataset consisted of Gram-positive and Gram-negative bacteria as examples of prokaryotic cells. For eukaryotic cells we looked at the entire dataset as well as the human subset of the eukaryotic data.

The sequence of the signal peptide and the first 30 amino acids of the mature protein from the secretory protein were used to construct positive examples. The first 70 amino acids of each cytoplasmic and for eukaryotes also nuclear proteins were used to construct negative examples of signal peptides.

The actual positive and negative samples were constructed by running a moving window of a particular size (21 amino acids in the case of eukaryotic cells and 17 amino acids in the case of prokaryotic cells). Each amino acid was encoded as a real number between $1 - 20$.

Table (1) states how many signal peptides and non-secretory proteins were used in the various datasets. Table (2) states how many positive (signal peptide) and negative samples (non-secretory proteins) this translates into after processing using the running window.

Table 1. Datasets used and number of sequences in datasets

Source	Signal peptides	Non-secretory proteins
Human	416	251
Eukaryote	1011	820
E. Coli	105	119
Gram-	266	186
Gram+	141	64

Table 2. The effective number of positive and negative examples after processing

Source	Signal peptides	Non-secretory proteins
Human	6293	10793
Eukaryote	14755	43460
Gram-	4541	9858
Gram+	3380	3392

3 Support Vector Machine Overview and Kernels Used

We are given ℓ examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)$, with $\mathbf{x}_i \in \mathcal{R}^n$ and $y_i \in \{-1, 1\}$ for all i. The problem of learning a function that will generalize well on new examples is ill-posed. The classical approach to restoring well-posedness to learning is regularization theory [9]. This leads to the following regularized learning problem:

$$\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda ||f||_K^2. \tag{1}$$

Here, $||f||_K^2$ is the norm in a Reproducing Kernel Hilbert Space $\mathcal{H}$ defined by a positive definite kernel function K, V is a *loss function* indicating the penalty we pay for guessing $f(\mathbf{x}_i)$ when the true value is y, and λ is a regularization parameter quantifying our willingness to trade off accuracy of classification for a function with small norm in the RKHS $\mathcal{H}$.

The classical SVM arises by considering the specific loss function

$$V(f(\mathbf{x}), y) \equiv (1 - yf(\mathbf{x}))_+, \tag{2}$$

where

$$(k)_+ \equiv \max(k, 0). \tag{3}$$

So the problem becomes:

$$\min_{f \in \mathcal{H}} \tfrac{1}{\ell} \textstyle\sum_{i=1}^{\ell} \xi_i + \lambda ||f||_K^2 \tag{4}$$

$$\text{subject to :} \quad y_i f(\mathbf{x}_i) \geq 1 - \xi_i \quad i = 1, \ldots, \ell \tag{5}$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, \ell. \tag{6}$$

Under quite general conditions it can be shown that the solution f^* to the above regularization problem has the form

$$f^*(\mathbf{x}) = \sum_{i=1}^{\ell} c_i K(\mathbf{x}, \mathbf{x}_i). \tag{7}$$

This can be written as the following dual program:

$$\max_{\alpha \in \mathcal{R}^\ell} \textstyle\sum_{i=1}^{\ell} \alpha_i - \alpha^T Q \alpha \tag{8}$$

$$\text{subject to :} \quad \textstyle\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{9}$$

$$0 \leq \alpha_i \leq C \qquad i = 1, \ldots, \ell, \tag{10}$$

where $C = \frac{1}{2\lambda\ell}$. Here, Q is the matrix defined by the relationship

$$Q = YKY \Longleftrightarrow Q_{ij} = y_i y_j K(x_i, x_j). \tag{11}$$

A geometric interpretation the RKHS norm

$$||f||_K^2 = \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(\mathbf{x}, \mathbf{x}_i),$$

is the margin M where $M = 1/2||f||_K^2$. For the case where the data can be perfectly separated (Figure 1) illustrates how minimizing the norm maximizes the margin.

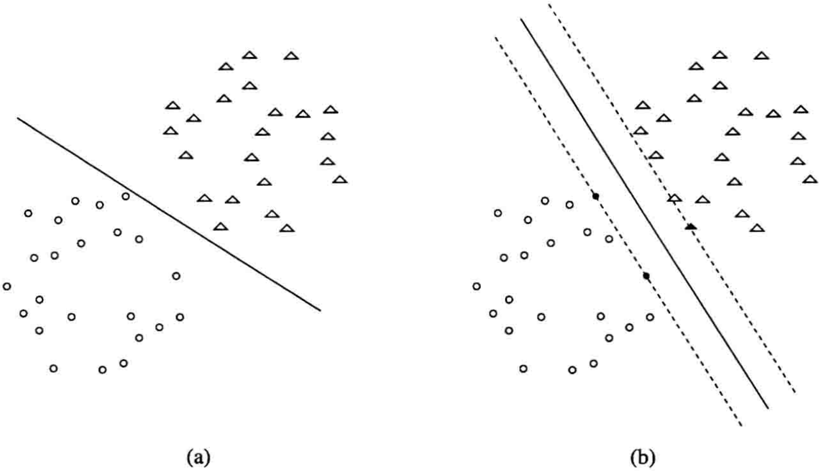

Fig. 1. Two hyperplanes with different margin. Intuitively, the large margin hyperplane (b) seems likely to perform better on future examples than the much smaller margin hyperplane (a)

We will use two types of kernels. One based upon Hamming distances and one based upon a similarity matrix, the Percent Accepted Mutation (PAM) Matrix. The input vectors $\mathbf{x}_i$ are not in $\mathcal{R}^n$ but are in the discrete space $\{1, ..., 20\}^n$. The following kernels based upon Hamming distances were used

$$K(\mathbf{x}_i, \mathbf{x}_j) = h(\mathbf{x}_i, \mathbf{x}_j) \tag{12}$$
$$K(\mathbf{x}_i, \mathbf{x}_j) = (h(\mathbf{x}_i, \mathbf{x}_j) + 1)^2, \tag{13}$$

where $h(\mathbf{x}_i, \mathbf{x}_j)$ is a count of how many elements in each position of the two sequences are identical.

The PAM matrix can be thought of as the probability that one amino acid replaces another so the similarity between two amino acids acids for example Leucine (L) and Serine (S)

$$K(L, S) = M_{LS} = P(\text{Leucine is replaced by Serine}).$$

We used the PAM250 matrix [7]. Note that this is not a valid kernel and also not a distance metric. However, we used this kernel anyway.

4 Results

We compare our classification results for the various datasets to results reported in [1,3] using neural networks, NNs, and Hidden Markov Models, HMMs. The results reported are 5 fold cross validation results (see tables (3 and (4). The designations SVM^1, SVM^2, and SVM^3 correspond to SVMs trained with the linear hamming distance, polynomial hamming distance and linear PAM250 matrix kernels. For most of the datasets the SVM results are at least as accurate as those of neural networks and HMMs. For the general eukaryotic our results are no better than that of the HMM, this is probably do to the fact that this is a large dataset and any reasonable algorithm will perform accurately. A cleavage site prediction defined as correct if the cleavage site falls anywhere in the sliding window. One interesting observation is that SVM^3 which is not using a valid kernel performs very well.

Table 3. Classification accuracy for SVMs, NNs, and HMMs for signal peptide versus non-secretory proteins

Algorithm	Eukaryotic	Human	Gram+	Gram-	E. coli
NN	97%	96%	96%	88%	89%
HMM	94%	-	96%	93%	-
SVM^1	96%	96%	97%	94%	91%
SVM^2	97%	97%	97%	94%	91%
SVM^3	98%	97%	97%	95%	92%

Table 4. Classification accuracy for SVMs, NNs, and HMMs for predicting cleavage sites

Algorithm	Eukaryotic	Human	Gram+	Gram-	E. coli
NN	70%	68%	68%	79%	84%
HMM	70%	-	65%	81%	-
SVM^1	73%	69%	82%	80%	83%
SVM^2	75%	72%	84%	79%	84%

5 Conclusions and Future Work

We are able to predict cleavage sites and also discriminate signal peptides from non-secretory peptides using a SVM classifiers. Our results at least as accurate as those using HMMs and NNs on the same task. It would be interesting to examine the support vectors selected in the training phase and analyze them as prototype signaling peptides and look at their statistical structure. It would also be of interest to apply a feature selection algorithm [8] to select which features/positions in the sequence are most relevant in making the above discriminations. Studying the above might yield some interesting biology. It would also be of interest to embed the HMMs used in these classification tasks into an SVM using the Fisher kernel [5]. An interesting note is that the kernel based upon the PAM250 matrix performed well even though it is not a valid kernel.

Acknowledgments

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR) Contract No. IIS-0112991, National Science Foundation (KDI) Contract No. DMS-9872936, and National Science Foundation Contract No. IIS-9800032.

Sayan Mukherjee would like to acknowledge the Sloan/DOE Fellowship in Computational Molecular Biology.

References

1. Nielsen, H., Brunak. S., von Heijne, G., Protein Engineering, vol. 12, no. 1, pp. 3-9, 1999.
2. Baldi, P., Brunak, S., Bioinformatics The Machine Learning Approach, M. I. T. Press, Cambridge, MA, 1999.
3. Nielsen, H., Engelbrecht, J., Brunak. S., von Heijne, G., Protein Engineering, vol. 10, no. 1, pp. 1-6, 1997.
4. V. Vapnik, Statistical Learning Theory, J. Wiley, 1998.
5. Jaakkola, T. and Haussler, D., Exploiting Generative Models in Discriminative Classifiers, NIPS 11, Morgan Kauffmann, 1998.

6. Bairoch, A. and Boeckmann, B., Nucleic Acids Research, 22, pp. 3578-3580.
7. Schwartz, R. M. and Dayhoff, M. O., Atlas of Protein Sequence and Structure, pp. 353-358, 1979.
8. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S. Choosing Many Kernel Parameters for Support Vector Machines, Machine Learning, 2002.
9. Tikhonov, A. N. and Arsenin, V. Y., Solutions of Ill-Posed Problems, W. H. Winston, Washington D. C., 1977.

Scaling Large Learning Problems with Hard Parallel Mixtures

Ronan Collobert[1,*], Yoshua Bengio[1], and Samy Bengio[2]

[1] Université de Montréal, DIRO
CP 6128, Succ. Centre-Ville, Montréal, Canada
{collober,bengioy}@iro.umontreal.ca
[2] IDIAP
CP 592, rue du Simplon 4, 1920 Martigny, Switzerland
bengio@idiap.ch

Abstract. A challenge for *statistical learning* is to deal with large data sets, e.g. in *data mining*. Popular learning algorithms such as *Support Vector Machines* have training time at least *quadratic* in the number of examples: they are hopeless to solve problems with a million examples. We propose a "hard parallelizable mixture" methodology which yields significantly reduced training time through modularization and parallelization: the training data is iteratively partitioned by a "gater" model in such a way that it becomes easy to learn an "expert" model separately in each region of the partition. A probabilistic extension and the use of a set of generative models allows representing the gater so that all pieces of the model are locally trained. For SVMs, time complexity appears empirically to locally grow *linearly* with the number of examples, while *generalization* performance can be enhanced. For the probabilistic version of the algorithm, the iterative algorithm provably goes down in a cost function that is an upper bound on the negative log-likelihood.

1 Introduction

As organizations collect more and more data, the interest in extracting useful information from these data sets with *data mining* algorithms is pushing much research effort toward the challenges that these data sets bring to statistical learning methods. One of these challenges is the sheer size of the data sets: many learning algorithms require training time that grows too fast with respect to the number of training examples. This is for example the case with Support Vector Machines [11] (SVM) and Gaussian processes [12], both being non-parametric learning methods that can be applied to classification, regression, and conditional probability estimation. Both require $O(T^3)$ training time (for T examples) in the worst case or with a poor implementation. Empirical computation time measurements on state-of-the-art SVM implementations show that training time grows much closer to $O(T^2)$ than $O(T^3)$ [2]. It has also been

[*] Part of this work has been done while Ronan Collobert was at IDIAP, CP 592, rue du Simplon 4, 1920 Martigny, Switzerland.

S.-W. Lee and A. Verri (Eds.): SVM 2002, LNCS 2388, pp. 8–23, 2002.