SUPPORT VECTOR MACHINES AS PATTERN CLASSIFICATOR IN BIOINFORMATICS

A. Madevska[1], D. Nikolic[2]

[1]Institute of Informatics, Faculty of Natural Sciences and Mathematics,
Sts. Cyril and Methodius University,
Arhimedova bb, PO BOX 162, Skopje, Macedonia
ana@ii.edu.mk
[2]Maastricht School of Management
Endepolsdomein, 150, 6229 EP. Maastricht, The Netherlands
nikolic@msm.nl

Abstract: We have explored Support Vector Machines (SVM) as pattern classificators in Bioinformatics. SVM has been applied to problems in Molecular Biology - the automatic classification of mouse chromosomes, regulatory sequences in a plant and predicting protein secondary structure. The results that we obtained are very promising compared to the results by other machine learning techniques.

Keywords: Support vector machines (SVM), optimization, classification, Bioinformatics, DNA sequences, chromosomes.

1. Introduction

Support vector machines (SVM) is a family of learning algorithms, which is currently considered as one of the most efficient method in many real world applications. The theory behind SVM was developed in the sixties and seventies by Vapnik and Chervonenkis, but the first practical implementation of SVM was only published in the early nineties. Since then the method gained more and more attention among the machine learning community thanks to its ability to outperform most other learning algorithms (including neural networks on decision trees) in many applications. As a result it has been successfully applied to all sorts of classifications issues, ranging from handwritten character recognition to speaker identification or face detection in images.

Recently SVM have been applied to biological issues, including gene expression data analysis or protein classification, particularly because of the high dimensionality of the data. As a result the research about SVM and computational biology is the object of much effort today, mainly due to researcher coming from the machine learning community. One can expect SVM to become a standard tool for bioinformaticians in the near future, just like clustering algorithms or dynamic programming methods today.

SVMs perform particularly well to the analysis of broad patterns of gene expression from DNA micro-array data. Micro-array technology allows putting a large number of DNA sequences under the influence of the same factor. This technology allows clustering genes in different groups according to the gene reaction under the same conditions. It is expected that some of these genes will share similar promoters.

Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information that can then be applied to gene-based drug discovery and development. The need for Bioinformatics capabilities has been triggered by the explosion of publicly available genomic information resulting from the Human Genome Project. The goal of this project - determination of the sequence of the entire human genome (approximately three billion base pairs) - will be reached by the year 2002. The science of Bioinformatics, which is the mixture of molecular biology with computer science, is essential to the use of genomic information in understanding human diseases and in the drug discovery.

2. SVM algorithm

The idea behind the SVM method is as follows: one maps N – dimensional vectors x of the input space X with appropriate Kernel function into high – dimensional (even infinite dimensional) vectors of the feature space.

Let x_i, i=1,..,N be a set of data points and y_i, i=1,.., N corresponding target classes.

The goal from the support vector learning is to get a classifier of the form

$$y = \mathrm{sign}(f(x)), \quad f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b \qquad (1)$$

where $k(x, x_i)$ is a Kernel function, in different forms:

- linear SVM: $k(x_i, x_j) = x_i^T x_j$;
- polynomial SVM: $k(x_i, x_j) = (1 + x_i^T x_j)^d$;
- Gaussian radial basis function SVM: $k(x_i,x_j) = \exp(-g \| x_i - x_j \|^2)$;

or other functions that satisfy Mercer's condition [3].

The sum in (1) goes over all training examples. The parameters α_i are determined by solving the following quadratic optimization problem:

$$\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i K_{i,j} \alpha_j - \sum_{i=1}^{N} \alpha_i \qquad (2)$$

$$\textit{where } K_{i,j} = y_i y_j k(x_i, x_j) \quad i, j = 1,\ldots,N$$

$$\textit{subject to } 0 \le \alpha_i \le C$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Training examples x_i for which

- $\alpha_i = 0$ are classified with $f(x_i) \geq 1$ if $y_i = 1$ and $f(x_i) \leq -1$ if $y_i = -1$;
- $0 < \alpha_i < C$ are classified with $f(x_i) = 1$ if $y_i = 1$ and $f(x_i) = -1$ if $y_i = -1$;
- $\alpha_i = C$ can be misclassified or are near boundary since with $f(x_i) \leq 1$ if $y_i = 1$ and $f(x_i) \geq -1$ if $y_i = -1$.

Patterns with $\alpha_i > 0$ are **support vectors**.

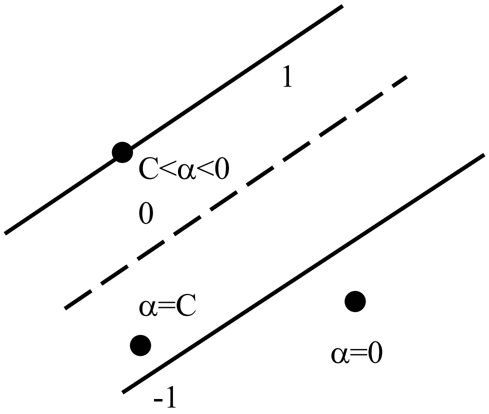

Figure 1: Support vectors

The classification function for SVM can be obtained by analyzing the SVM for the linear case and then generalized to the nonlinear case. Finding the support vectors (in linear case) is based on maximal margin separating hyperplanes. First it is assumed that the two classes are separated by two parallel hyperplanes

$w^T x + b = 1$ and $w^T x + b = -1$

$$\begin{cases} w^T x_i + b \geq 1 & \text{if} \quad y_i = 1 \\ w^T x_i + b \leq -1 & \text{if} \quad y_i = -1 \end{cases} ,i = 1,\ldots,N \qquad (3)$$

The distance between these two hyperplanes is $2/\|w\|$.

The generalization performance is improved (Statistical Learning Theory) by maximizing the distance, or by solving the dual problem (minimizing $w^T w$):

$$\begin{aligned} &\min_{w} w^T w \\ &\text{subject to } y_i(w^T x_i + b) - 1 \geq 0 \end{aligned} \tag{4}$$

The constraint is the same as in (3), but rewritten in one equation.

In the case of nonseparable data, where misclassification is possible, the constraints are relaxed by introducing nonnegative, slack variables ξ_i. The presence of a positive ξ_i is penalized by an extra term $C\sum_{i=1i-}^{N} \xi_i$ in the objective function.

$$\begin{aligned} &\min_{w} w^T w + C\sum_{i=1}^{N} \xi_i \\ &\text{subject to } \begin{cases} y_i(w^T x_i + b) - 1 + \xi_i \geq 0 \\ \qquad\qquad \xi_i \geq 0 \end{cases} \end{aligned} \tag{5}$$

The dual problem:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i K_{i,j} \alpha_j - \sum_{i=1}^{N} \alpha_i \\ &\text{where } K_{i,j} = y_i y_j x_i{}^T x_j \quad i,j = 1,...,N \\ &\text{subject to } 0 \leq \alpha_i \leq C \\ &\qquad \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned} \tag{6}$$

where the variables from the equation (1) are

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{7}$$

and b is recovered by the expression

$$b = y_j - \sum_{i=1}^{N} y_i \alpha_i x_i^T x_j \text{, over all patterns j for which } 0 < \alpha_j < C. \tag{8}$$

Substituting (7) in $f(x)=w^T x + b$, one gets:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i x_i^T x + b \tag{9}$$

which is the SVM classifier in the non-separable case.

3. SVM in Bioinformatics

Mouse Chromosome classification

Mouse chromosomes are more difficult to classify than human chromosomes, so the possibility of automatic classification becomes even more important. The database consists of 3723 mouse chromosomes. The inputs to the SVM's are vectors of 30 discrete values, representing the banding profile and the chromosome's length. There are 21 different chromosome classes. The solution to the classification problem is achieved by constructing 21 decision hyperplanes. Classification is done by the highest value of the "one against the others" separators.

For the classification task, different Kernel functions are used. The training is done mostly with Gaussian Radial Basis Function (GRBF) as a Kernel function for the SVM, but the polynomial Kernel functions are considered too.

Most of the results are obtained for the training set of 2250 chromosomes versus test set of 1473 mouse chromosomes. The partition of the data set is the same as in [1], so the results can be compared.

Results: The experiments that gave the best results are the ones with the Gaussian Radial Basis Function used as a Kernel function. The results using Polynomial function are also given.

3.1 Gaussian Radial Basis Function

g , gamma in $K(x_i,x_j) = \exp(-\text{gamma} \| x_i-x_j \|^2)$; **C,** trade-off between training error and margin;

Percentage of correctly recognized test patterns for different values of gamma in GRFB:

C	100	**100**	100	100	100
g	0.005	**0.1**	0.12	1	10
percentage	77.8	**87.2369**	87.033	84.5893	82.4168

Table 1

Since the best results were produced for gamma value of 0.1, further experiments are carried out toward finding the right value of C in GRBF, considered the mentioned gamma value.

Percentage of correctly recognized test patterns for different values of C in GRBF:

g	0.1	0.1	**0.1**	**0.1**	0.1	0.1	0.1
C	50	105	**110**	**115**	120	125	150
percentage	86.6938	87.3048	**87.3727**	**87.3727**	87.3048	87.169	87.033

Table 2

For g=0.1, better results are achieved for C value of 110, or 115.

3.2 Polynomial function

Table 3 shows the percentage of correctly recognized test patterns for different values of the degree d in the Polynomial Kernel function, where the trade off between the margin and the training error is 1000.

d	2	5	10	15
percentage	80.5838	83.8425	83.8425	**84.5893**

Table 3

Further experiments are carried out for different number of training vs. test patterns.

Because of the results achieved with GRBF SVM, further experiments have values g=0.1, C=110. Training sets are chosen to have the same number of representatives for each class (except for the last one that has only 19 examples). The results are given in the following table.

Training/test	2019 / 1704	2220 / 1503	**2420 / 1303**	3000 / 723
Percentage	85.622	87.6248	**87.9509**	87.4136

Table 4

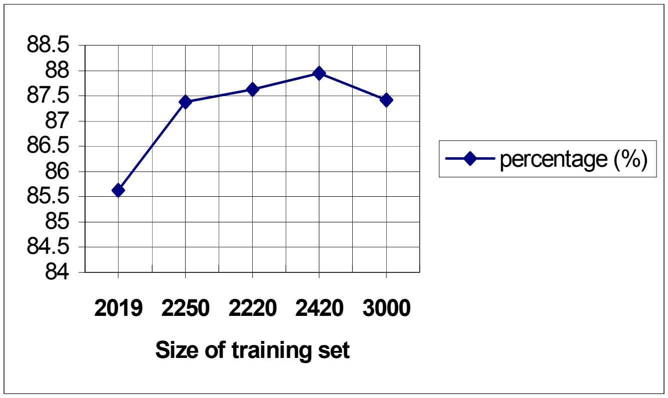

Figure 2: Percentage of correctly recognized test set for the different training/test partitions of the data set

Summary: The best result **87.96%** was obtained for GRBF, with

- gamma value 0.1 (g = 0.1);
- train / test set : 2420 / 1303;
- difference between training error and margin C = 110;

The best published result over the same database (partition 2250 / 1473) is accomplished by using Radial Basis Function Neural Networks is 87.3 %.

For the same training / test set (2250 / 1473), the SVM's result is 87.4 % The comparison of the published results on classifying mouse chromosomes using RBF, and the SVM method over the identical data, shows slightly better results in favor of SVM.

Regulatory sequences in Arabidopsis Thaliana

This work has been done in cooperation with the data mining group at the SISTA - ESAT, KU Leuven. They have used the data – DNA sequences of *Arabidopsis Thaliana* from the Bioinformatics team of the Genetics Department of the Flanders Institute of Biotechnology at the University of Ghent.

Arabidopsis Thaliana has shortest genome among plants. Its regulatory sequences (genetics switches) are short (5 – 10 base pairs). Each gene has a specific pattern of regulatory sequences. They are placed 40 – 1000 bp upstream of a gene. Repressor and activator proteins bind on both strands.

Used data are manually annotated and are surely correct

The trainings set are specified for 1 specific element, G-box, which has the most entries in the database PlantCARE (18 sequences and sites).

Results: The recognition process using ansamble of 75 MLP gave poorer results than models build by SVM. The final results will be obtained when further improved experiments using SVM will be undertaken, cocidering the fact that there are a lot more negative examples than positive ones.

Predicting the secondary structure of globular proteins

Results: The published results for this problem [2] are 64% using MLP. In this case the SVM gives poor results for the same training/test set, only 62%. It is not clear if the training / test sets were the same as with the Qian, Sejnowski.

4. Conclusion

Support vector machines were used on a pattern classification problems, in the domain of Molecular Biology. The main advantage of SVMs is that they can easily deal with a large number of features (thousands of genes) and a small number of training patterns (dozens of patients).

SVM appear to be the state-of-the-art methods in many other applications, so it is reasonable to think that this relatively new technology will generate interesting progress in the coming years in bioinformatics too.

5. Acknowledgements

The authors wish to express their gratitude and acknowledgement to M.T. Musavi, University of Maine, Orono, USA, for allowing them to make use of the mouse chromosome database (University's of Maine public ftp site).

6. References

1. M.T.Musavi et.al, "Mouse chromosome classification by radial basis function network with fast orthogonal search", Neural Networks 11 (1998) 769-777.
2. Qian, Sejnowski "Predicting the secondary structure of globular proteins", JMB. 1988;
3. A. Madevska, D. Nikolic, "Automatic Classification With Support Vector Machines In Molecular Biology", *Proceedings of III International Conference on Cognitive and Neural Systems*, Boston, MA, USA, 1999;
4. K.J.C. Burges, "A tutorial on Support Vector Machines for Pattern Recognition" , *Data mining and knowledge discovery,* 1998;
5. V.N. Vapnik, The nature of statistical learning theory, Springer, 1995;
6. C. Bishop, Neural Networks for Pattern Recognition, Oxford Press, 1998;
7. Jean-Philippe Vert, "Introduction to support vector machines and applications to computational biology ", DRAFT ,July 17, 2001