SUPPORT VECTOR MACHINES ON HEMOGLOBIN SECONDARY STRUCTURES

Turgay İbrikci*, Ayça Çakmak*, İrem Ersoz**, Okan K. Ersoy***

*Cukurova University Dept. of Electrical- Electronics Engineering, Adana, Turkey
**Mersin University Faculty of Technical Education Dept. of Electronics Education, Mersin, Turkey
**Purdue University School of Electrical and Computer EngineeringWest Lafayette, Indiana, USA

ibrikci@cukurova.edu.tr

Abstract: Secondary structure prediction of proteins has increasingly been a central research area in bioinformatics. In this paper, support vector machines (SVM) are investigated in terms of accuracy of prediction of hemoglobin secondary structures. For this purpose, the training and testing data are obtained from the Protein Data Bank, USA with DSSP structures. The results of prediction with window size of 17 were found to be 98.62 % for helix, 75.38 % for coil and overall 92.26% when the using Gaussian radial basis kernel. The Matthew's corelation coefficent factors are 0.8 for both predictions.

Introduction

The protein molecules represent much of the bulk of an organism and accomplish almost all of its biochemical activities. To understand the life process of an organism, it is necessary to know the protein's structure since it is closely related to its function. The major function of the hemoglobin is to collect oxygen that diffuses into the plasma of the blood from the lungs, then to deliver it through the arteries to the tissues for maintaining the viability of cells, and to transport carbon dioxide back to the lungs through the veins [1]. All adult hemoglobin throughout the world has the same structure. However, sometimes some defects can occur in the genetic code for hemoglobin, and cause abnormalities. The types of disorders that can result include sickle cell disease and thalassemia [2].

Methods of predicting protein structure have been improved in the late 1990s through the use of statistical and computational learning methods, starting with the Krighbaum-Kuntton and Chou_Fasman methods [3][4]. Krighbaum-Kuntton used the multiple linear regression algorithms to predict the amino acid composition of a protein [3]. This attempt continued with the Chou-Fasman method that achieved a three-state (Q_3) accuracy of 52% [4]. This method is a well-known empirical statistical algorithm that is based on the frequencies of the secondary structure types. Within the field of protein secondary structure prediction, the idea of combining different prediction methods is also well-established.

Support vector machines have been applied to bioinformatics problems; one of them is secondary structure prediction [5], [6]. Hua and Sun in 2001 used SVMs in a secondary structure prediction scheme. They used only residue information, with an acccuracy of 73.3 % for Q [7]. Yu-Dong Cai, Xiao-Jun Liu, Xue-biao Xu, and Kuo-Chou used the sliding windows technique with SVM to test a set of protein sequences based on group classification learned from a training set. Their prediction accuracy was 75.2% for three-state (Q_3) [8]. The prediction of secondary structure is the first step for prediction of protein tertiary structure. SVMpsi was developed by Hyunson Kim and Haesun Park in 2003 to improve the current level of prediction by incorporating new tertiary classifiers and their jury decision system [9]. They achieved different Q_3 values on different datasets. The maximum accuracy was 81.8 % on the SOV94, which is a non-homologues dataset. Ward, McGuffin, Buxton, and Jones applied binary SVM with polynomial kernel to proteins. The average three-state (Q_3) prediction accuracy was 77.07 ± 0.26 % on the 121 non-homologues proteins [10].

Datasets and Methods

In this study, the dataset was obtained from the Protein Data Bank, USA and consisted of 13250 hemoglobin chains [11]. The DSSP assignments, which are defined as the secondary structure in eight categories, α - helix, 3_{10} helix, π helix , extended strand, isolated β Bridge, turn, bend, rest or coil were used [12]. The SVM was trained to predict the three categories which are helix, strand and coil. The training files contain a primary structure and its corresponding secondary structure. These structures were investigated with a number of sizes of sliding windows that consist of contiguous amino acid residues.

In this study, the size of window used was 17. The window size is chosen according to the following formula with *r* number of amino acid residues before and after the center element of the window:

$$\text{window size (W)} = 2*r+1$$

ISSN: 1727-1983

The centering technique is based on the assumption that the central amino acid has a large influence in the structural classification of that window [13]. Prediction is applied by labeling the input pattern with the secondary structure.

Support Vector Machines

Support vector machines (SVMs) constitute a supervised learning algorithm, and was first discussed by Vapnik in the 1960s for the two-class classification problem [13]. The SVM is a training algorithm for learning classification rules, and uses a hypothesis space of linear functions in a high dimensional feature space, and incorporates latest advances in optimization theory as applied to statistical learning theory [14]. Two key elements of SVM are the techniques of mathematical programming and kernel functions, which are needed for mapping the input vectors to high-dimensional feature vectors [15]. Some candidate kernel functions are linear, polynomial, sigmoid function and radial basis function (RBF).

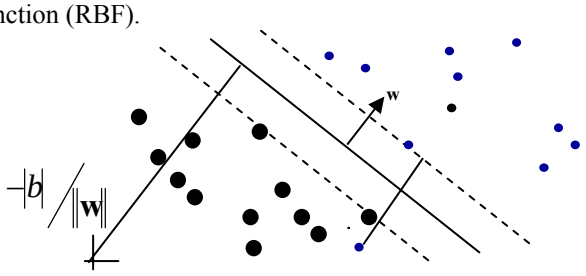

Figure1: The visualization of 2-D classification space with relevant parameters.

In the feature space, the hyperplane for linear classification is defined by

$$w^T \cdot x_i + b \geq +1 for all \Rightarrow y_i \in +1$$
$$w^T \cdot x_i + b \leq -1 for all \Rightarrow y_i \in -1$$

where the terms used are explained in Figure 1.

The decision rule is given by

$$f_{w,b}(x) = \mathrm{sgn}(w^T x + b)$$

SVMs are used to find nonlinear separating surfaces by using kernel functions which aid in transforming input vectors in to feature vectors nonlinearly. The length of feature vectors can be very long since learning can be done in the dual space where the complexity of computations is linearly related to the size of the training dataset, and not to the length of feature vectors. In optimization theory, it is known that a quadratic programming problem has an equivalent dual problem that is sometimes more tractable than the original problem. The optimization equations in the dual space are given by

$$\min_{\alpha} \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^{m} \alpha_i$$

$$such\ that \sum_{i=1}^{m} y_i \alpha_i = 0$$
$$0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, m$$

The vector α is referred to as a dual space vector variable, and replaces w and b in the original formulation. K ($x_i.x_j$) function is called the kernel function.

Kernel functions are usually chosen as nonlinear for mapping input vectors in to feature vectors. We assume that the training set is given by

$$S = \{(x_1,y_1), (x_1,y_1), \ldots. (x_m,y_m), \}$$

being the set of labeled examples. A training vector is given by

$$\vec{x} = (x_1, x_2, x_3, \ldots. x_m)$$

Radial basis function kernels can be written as

$$K(\vec{x}i, \vec{x}j) = \exp\left(-\frac{1}{2\alpha^2}\left\|\vec{x}i - \vec{x}j\right\|^2\right)$$

where α is a parameter which controls the Gaussian width of the kernel; α is usually set equal to the median of the Euclidean distances from each positive example to the nearest negative example. The output is dependent on the Euclidean distance between x_j and x_i.

Experiments

The SVM with kernel Gaussian-RBF was tested using hemoglobin data sets with the three-class assignments of the DSSP. The dataset is consisted of hemoglobin with 13250 amino acid-chains. This sequence was used with sliding window size 17 for the prediction of the secondary structure of hemoglobin. This window size is optimum window size of the study [16]. leave-one-out validation was used for validating the results. The average prediction rate was calculated, providing a way of evaluating the performance of SVM trained with kernel Gaussian-RBF. The quality of prediction for DSSP assignments are denoted by overall (Q_{total}), correct classification (CC), sensitivity (SE), and Matthew's (also called Pearson) correlation coefficient (C) to validate the results. They are computed with the following formulas:

$$Q_{total} = 100 * \frac{TP}{N}$$

$$CC = 100 * \frac{TP + TN}{N}$$

$$SE = 100 * \frac{TP}{TP + FN}$$

$$C = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

ISSN: 1727-1983

where

N : the total number of predicted residues,
TP : the number of true positives,
TN : the number of true negatives,
FP : the number of false positives,
FN : the number of false negatives.

Sensitivity and specificity give a measure of how well true positive patterns and true negative patterns are correctly identified, respectively. The correlation coefficient is always between -1 and +1, with +1 showing complete agreement, -1 complete disagreement and 0 showing that the prediction was uncorrelated with the results.

Results and Discussion

The results are given with Q, correct classification (CC), sensitivity (Se), and Pearson correlation coefficients(C) in Table 1.

Table 1: The results of the study

	Q (%)	CC (%)	Se (%)	C
Helix	98.62	92.26	77.62	0.8
Coil	75.38	92.26	22.37	0.8
Overall	92.26	91.39	50	0.8

The results are comparable with other studies. Irem Ersoz and Turgay Ibrikci etc. used generalized regression neural network (GRNN), probabilistic neural network (PNN) and backpropagation network (BPN) algrithms on the same hemoglobin data set[16]. The final results they achieved were Q_{GRNN} = 90.04%, Q_{PNN} = 90 %, and Q_{BP} = 87.18 % .

Hua and Sun used SVMs in secondary prediction scheme, their Q was 73.5 % [7]. Hyunson Kim and Haesun Park achieved Q = 81.8 % on non homologues protein sets.

When the results of this study were compared with their results, they are extremely good. This study results were better 2.22 % than GRNN, 2.26 % than PNN, and %5.08 than BPN on the same data set. Thus, the SVM was successful on these small data sets, and the chosen kernel function Gaussian-RBF was founded to be more appropriate for this data.

Acknowledgment

This research has been supported by Cukurova University Research Foundation MMF2004-BAP21 and TUBITAK-EEEAG-100E037. The authors would like to thank Prof. Dr. Seyhan Tukel, our colleague at Cukurova University for her help in understanding the hemoglobin structures and useful discussions.

References

[1] RINSHO N. (1998) : ' Human Hemoglobin Structure and Respiratory Transport. Department of Human Genetics', *National Institute of Genetics*, 9,(54), pp. 2320-2325

[2] PAULING L.(1952): 'The Hemoglobin Molecule in Health and Disease', *Proceedings of the American Philosophical Society*, 96, No: 5, pp. 555-565

[3] CAI Y., LIU X., XU X., CHOU K. (2002): 'Artificial Neural Network for Predicting Protein Secondary Structure Content', *Computers and Chemistry*, 26, pp. 347-350

[4] CHOU P.Y., FASMAN G.D.(1978): 'Empirical Predictions of Protein Conformation', *Annual Review Biochemistry*, 47, pp. 251-276

[5] CAI Y., LIU X., XU X., ZHOU G.(2001): 'Support Vector Machines for Predicting Protein Structural Class', *Bioinformatics*; 2, no.3 www.biomedcentral.com/1471-2105/2/3

[6] GUERMEUR Y.(2002): 'Combining Discriminant Models with New Multi-Class SVMs', *Pattern Analysis and Applications*, 5(2), pp.168-179, 2002.

[7] HUA S., SUN Z. (2001): 'A novel method of protein secondary structure prediction with high segment overlap : Support vector machine approach', *Jour. Mol. Biol.*, ,308, pp: 397-407

[8] CAI Y, LIU X., XU X., CHOU K. (2000): 'Support Vector Machines for Prediction of Protein Subcellular Location', *Molecular Cell Biology Research Communication*, 4, pp:230-233

[9] KIM H., PARK H.(2003): Technical report https://wwws.cs.umn.edu/tech_reports_upload/tr2003/03-005.pdf

[10] WARD J.J., MCGUFFIN B.F., BUXTON D.T. (2003): 'Secondary structure prediction with support vector machines', *Bioinformatics,* 19 (13) pp:1650-1655

[11] PROTEIN DATA BANK, Brookhaven National Laboratory : www.rcsb.org/pdb

[12] KABSCH W., SANDER C. (1983): 'Dictionary of Protein Secondary Structure: Pattern recognition of hydrogen-bonded and geometrical features', *Biopolymers*, 22(12), pp. 2577-637

[13] QIAN N., SEJNOWSKI T.J. (1988): 'Predicting the Secondary Structure of Globular Proteins Using Neural Network Models', *J. Mol. Biol.*, 202, pp. 865-884

[14] VAPNIK V. N.(1995): 'The Nature of Statistical Learning Theory', New York: Springer

[15] CRITIANINI J. TAYOR S. (2000) : 'An Introduction to Support Vector Machines', Cambridge: Cambridge University Press

[16] ERSOZ, I., IBRIKCI T., CAKMAK A., ERSOY O., (2005): 'Secondary structure prediction of hemoglobin by using neural network methods', *(In review).*

ISSN: 1727-1983