HEMOGLOBIN SECONDARY STRUCTURE PREDICTIO WITH FOUR KERNELS ON SUPPORT VECTOR MACHINES

T. Ibrikci[1], A. Cakmak[1], I. Ersoz[2] , O. K. Ersoy[3]
[1] Department of Electrical-Electronics Engineering, Çukurova University, Adana, Turkey
[2] Technical Education Department, Mersin University, Tarsus, Turkey
[3]Purdue University School of Electrical and Computer Engineering, West Lafayette, Indiana, USA

Abstract - **Secondary structure prediction of proteins has increasingly been a central research area in bioinformatics. In this paper, support vector machines (SVM) are discussed as a method for the prediction of hemoglobin secondary structures. Different sliding window sizes and different kernels of SVM are comparatively investigated in terms of accuracy of prediction of hemoglobin secondary structure. For this purpose, the training and testing data were obtained from the Protein Data Bank, US with Database of Secondary Structures of Protein (DSSP). The results of prediction with different SVM kernels and different window sizes were found to be in the range of 5.93 - 15.90 , 67.76-70.05 , 69.77 - 73.25 , and 74.42 - 77.64 % for linear kernel, sigmoid kernel, polynomial kernel and Gaussian radial basis kernel, respectively.**

I. 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 being secondary structure prediction [5], [6]. 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) [7]. 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 [8]. They archived 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 [9].

II. Dataset And Methology

In this study, the dataset was obtained from the Protein Data Bank, USA and consisted of 2820 hemoglobin chains [10]. The DSSP assignments, which are defined as the secondary structure into eight categories, α - helix, 3_{10} helix, π helix , extended strand, isolated β Bridge, turn, bend, rest or coil were used [11]. The networks were trained to predict these categories with k-fold cross validation. 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 eight different sizes of windows used were 11-13-15-17-19-21-23-25. Each 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:

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

The centering technique is based on the assumption that the central amino acid has a large influence in the structural classification of that window [12]. Prediction is applied by labeling the input pattern with the secondary structure.

III. Support Vector Machines

The support vector machines (SVMs) constitute a supervised learning algorithm, and was first discussed by Vapnik in the 1960s for the two-class classification problem

1-4244-0020-1/05/$20.00 â2005 IEEE

[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. Some candidate kernel functions are linear, polynomial, sigmoid function and radial basis function (RBF).

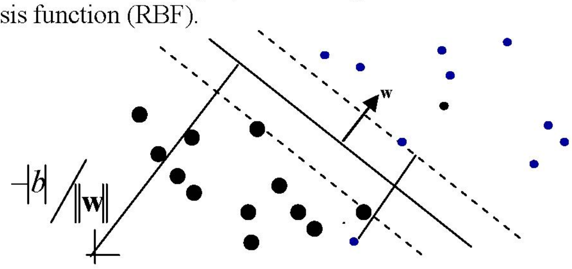

Fig.1. The visualization of 2-D classification space with relevant parameters.

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

$$w^T \times x_i + b \text{ ³ } +1 \text{forall } \text{§ } y_i \text{ Î } +1$$
$$w^T \times x_i + b \text{ £ } -1 \text{forall } \text{§ } y_i \text{ Î } -1$$

where the terms used are explained in Figure 1.

The decision rule is given by

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

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

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

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

The vector □ 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.

IV. Kernels

Kernel functions are usually chosen as nonlinear for mapping input vectors in to feature vectors. Below we discuss some significant kernel functions. 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

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

A. Linear Kernel

The linear kernel is given by

$$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \bullet \vec{x}_j$$

B. Sigmoid Kernel

Sigmoid kernel is given by

$$K(\vec{x}_i, \vec{x}_j) = \tanh(s * \vec{x}_i \bullet \vec{x}_j + r)$$

where *s* is called gain, and *r* is a coefficient factor which is called threshold.

C. Polynomial Kernel

The polynomial kernel is given by

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \bullet \vec{x}_j + 1)^d$$

where d is the degree of kernel. This kernel is equivalent to mapping the original data space into a higher order polynomial vector space, and finding a linear separating surface in that space.

D. Radial Basis Function Kernel

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.

IV. Experiments

The hemoglobin datasets were used with eight different window sizes to obtain comparative accuracy rates of secondary structure prediction for eight classes. K-fold validation was used for validating the results. K-fold validation randomly divides the training dataset into k sets. SVM is trained k times, each time leaving one set out as the testing set. The average prediction rate is calculated, providing a way of evaluating the performance of SVM trained with different kernels. The accuracy rates for DSSP assignments are denoted by Q_{total}. These results were then averaged to determine overall accuracy rates in which the accuracy Q_{total} was computed with the following formula:

$$Q_{total} = \frac{\text{CorrectClasses}}{N} \%$$

where *N* shows the total number of predicted residues. The DSSP class assignments on the hemoglobin dataset were α - helix, 3_{10} helix, π helix, extended strand, isolated β bridge, turn, bend, rest or coil, and average accuracy rate.

V. RESULTS AND DISCUSSION

SVM kernels were tested (linear, sigmoid, polynomial and Gaussian-RBF) using hemoglobin data sets with the eight class assignments of the DSSP. The dataset consisted of hemoglobin with 2820 amino acid-chains. This sequence was used with sliding windows for the prediction of the secondary structure of hemoglobin.

Each window size was applied to different kernels in SVM with hundred-fold validation. For this purpose, the whole data set was randomly divided into 100 subnets of equal size. There was observed a big accuracy variation among the kernels. Especially with the linear kernels, this ranged between 5.93 to 15.90 %, which are smaller than the values of sigmoid, polynomial, and radial basis kernels. The accuracy values with the radial basis kernel were higher and ranged between 74.42 to 77.64 %. However, variability with respect to the window sizes was smaller. The highest accuracy was with the window size 11 and Gaussian-RBF kernel whereas the lowest average was with window size 19 and linear kernel. It is obvious that the linear kernel percentages are significantly less than other kernels. Thus, the linear kernel is not suitable for predicting secondary structures of protein. This also indicates that the secondary structure prediction is a nonlinear classification problem.

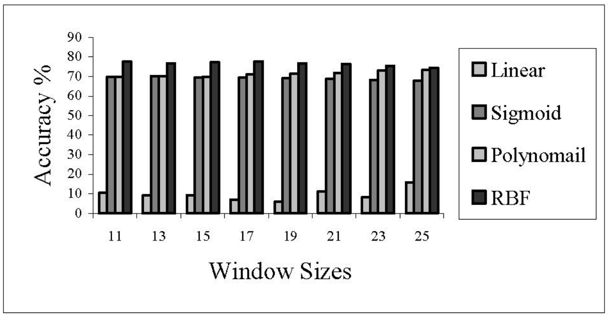

Fig. 2. The accuracy (%) of kernels with window sizes.

The average hemoglobin prediction values are shown in Figure 3, ranging between 9.59 to 76.43 % with four kernels. The highest percentage of prediction accuracy was recorded as 76.43 % with the Gaussian-RBF kernel second highest percentage was with the polynomial kernel of degree 3.

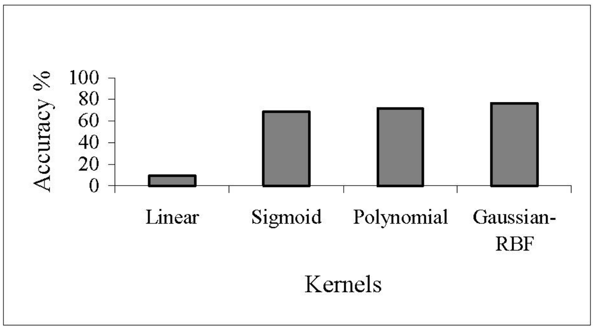

Fig. 3. The average prediction values

VI. CONCLUSION

The experimental findings indicate that sigmoid kernel, polynomial kernel, and Gaussian Radial Basis kernel are the applicable kernels in hemoglobin secondary structure prediction. Among these, the radial basis function kernel gave the highest accuracy rates in the experiments. Dependence on the sliding window size was less significant, but the best window size was approximately 11.

ACKNOWLEDGMENT

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

REFERENCES

[1] N.Rinsho, "Human Hemoglobin Structure and Respiratory Transport. Department of Human Genetics," National Institute of Genetics, vol.9, 54, pp. 2320-2325, 1998.

[2] L.Pauling, "The Hemoglobin Molecule in Health and Disease," Proceedings of the American Philosophical Society, vol. 96, No: 5, pp. 555-565, October, 1952.

[3] Y.Cai, X. Liu, X. Xu, K. Chou, "Artificial Neural Network for Predicting Protein Secondary Structure Content," Computers and Chemistry, vol. 26, pp. 347-350, 2002.

[4] P.Y Chou, G.D. Fasman, "Empirical Predictions of Protein Conformation", Annual Review Biochemistry, vol. 47, pp. 251-276, 1978.

[5] Y. Cai, X. Liu, X., X. Xu, G. Zhou, "Support Vector Machines for Predicting Protein Structural Class," Bioinformatics; vol. 2, no.3 www.biomedcentral.com/1471-2105/2/3 (2001)

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

[7] Yu-Dong Cai, Xiao-Jun Liu, Xue-biao Xu, Kuo-Chou, "Support Vector Machines for Prediction of Protein Subcellular Location," Molecular Cell Biology Research Communication, vol. 4, pp:230-233, 2000.

[8] Hyunsoo Kim and Haesun Park, 2003 Technical report https://wwws.cs.umn.edu/tech_reports_upload/tr2003/03-005.pdf

[9] J.J. Ward, L.J. McGuffin, B.F. Buxton, D.T. Jones, "Secondary structure prediction with support vector machines," Bioinformatics vol:19 no:13 pp:1650-1655, 2003.

[10] PROTEIN DATA BANK, Chemistry Department, Brookhaven National Laboratory. Upton, NY 11973 USA, www.rcsb.org/pdb/

[11] W. Kabsch, C. Sander, "Dictionary of Protein Secondary Structure: Pattern recognition of hydrogen-bonded and geometrical features," Biopolymers, vol. 22: 12, pp. 2577-637, 1983.

[12] N. Qian, T.J. Sejnowski, "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models," J. Mol. Biol., vol.202, pp. 865-884, 1988.

[13] V. N. Vapnik, The Nature of Statistical Learning Theory. New York: Springer. 1995.

[14] N. Critianini, J. Shawe-Taylor, An Introduction to Support Vector Machines. Cambridge: Cambridge University Press. 2000.