

Multi-Class Support Vector Machines for Protein Secondary Structure Prediction

Minh N. Nguyen

minhnguyen@pmail.ntu.edu.sg

Jagath C. Rajapakse

asjagath@ntu.edu.sg

School of Computer Engineering, Nanyang Technological University, Singapore

Abstract

The solution of binary classification problems using the Support Vector Machine (SVM) method has been well developed. Though multi-class classification is typically solved by combining several binary classifiers, recently, several multi-class methods that consider all classes at once have been proposed. However, these methods require resolving a much larger optimization problem and are applicable to small datasets. Three methods based on binary classifications: one-against-all (OAA), one-against-one (OAO), and directed acyclic graph (DAG), and two approaches for multi-class problem by solving one single optimization problem, are implemented to predict protein secondary structure. Our experiments indicate that multi-class SVM methods are more suitable for protein secondary structure (PSS) prediction than the other methods, including binary SVMs, because their capacity to solve an optimization problem in one step. Furthermore, in this paper, we argue that it is feasible to extend the prediction accuracy by adding a second-stage multi-class SVM to capture the contextual information among secondary structural elements and thereby further improving the accuracies. We demonstrate that two-stage SVMs perform better than single-stage SVM techniques for PSS prediction using two datasets and report a maximum accuracy of 79.5%.

Keywords: bioinformatics, multi-layer perceptrons, protein structure, secondary structure prediction, support vector machines (SVMs), multi-class SVMs

1 Introduction

Proteins are large biological molecules with complex structures and constitute to the bulk of living organisms: enzymes, hormones and structural material. The function of a protein molecule in a given environment is determined by its 3-dimensional (3-D) structure [1]. Protein 3-D structure prediction directly from amino acid sequences still remains as an open and important problem in life sciences. The bioinformatics approach first predicts the protein secondary structure (PSS) which represents an 1-D projection of the very complicated 3-D structure of a protein [12]. The goal of secondary structure prediction is to classify a pattern of residues in amino acid sequences to a corresponding secondary structure element: an α -helix (H), β -strand (E) or coil (C, the remaining type).

Many computational techniques have been proposed in the literature to solve the PSS prediction problem, which can be broadly fallen into three categories: (1) statistical methods, (2) neural network approaches, and (3) nearest neighbor methods. The statistical methods are mostly based on likelihood techniques [4, 5, 6]. Neural network approaches use residues in a local neighborhood or window to predict the secondary structure at a particular location of an amino acid sequence [9, 15]. The nearest neighbor method often uses the k -nearest neighbor techniques [16, 17]. SVMs have been earlier applied to PSS prediction [8]. One of the drawbacks in this approach is that the method does not capture the global information of the amino acid sequence due to the limited size of the local neighborhood. Additionally, the method only constructs a multi-class classifier by combining several binary classifiers.

Despite the existence of many approaches, the current success rates of existing approaches are insufficient; further improvement of the accuracy is necessary. Most existing secondary structure techniques are single-stage approaches, except the PHD [15] and PSIPRED [9] methods which combined two multi-layer perceptron (MLP) networks. Single-stage approaches are unable to find complex relations (correlations) among different elements in the sequence. This could be improved by incorporating the interactions or contextual information among the elements of the output sequence of secondary structures. We argue that it is feasible to enhance present single-stage approaches by augmenting with another prediction scheme at their outputs and propose to use SVMs as the second-stage.

This paper investigates the use of multi-class support vector machines for PSS prediction. We present two multi-class SVM techniques with three methods based on binary classifiers to PSS prediction. And then, we cascade two multi-class SVMs for the prediction scheme to improve the prediction accuracy from the output of the first stage. We report a prediction accuracy of 79.5% with PSIPRED dataset [9] based on PSI-BLAST profiles.

2 Single Stage SVM Approaches

Let us denote the given amino acid sequence by $\mathbf{r} = (r_1, r_2, \dots, r_n)$ where $r_i \in \Sigma_R$ and Σ_R is the set of 20 amino acid residues, and $\mathbf{t} = (t_1, t_2, \dots, t_n)$ denotes the corresponding secondary structure sequence where $t_i \in \Sigma_T$ and $\Sigma_T = \{H, E, C\}$; n is the length of the sequence. The prediction of the PSS sequence, \mathbf{t} , from an amino acid sequence, \mathbf{r} , is the problem of finding the required mapping from the space of Σ_R^n to the space of Σ_T^n .

As in most existing methods, the input domain, we consider, is made up of 21 input positions: one for each amino acid and one for a padding space to indicate overlapping end of a sequence. Only one of the 21 components has a value of 1 for given input residue while the rest of the components are 0. When padding for the end of the sequence is required, the padding input component is set to 1. Let v_i be the orthogonal binary vector representing 21-dimensional coding of the residue and the input pattern to the SVM to predict the residue at site i be $\mathbf{v}_i = (v_{i-h_1}, v_{i-h_1+1}, \dots, v_i, \dots, v_{i+h_2})$ where v_i denote the center element, h_1 and h_2 denote the width of window on two sides, and $w_1 = h_1 + h_2 + 1$ is a neighborhood size around the element i . For PSI-BLAST profiles, v_i are the values from raw profile matrices scaled to the $[0, 1]$ range [9].

In what follows in this section, we present several multi-class support vector machines (SVM) for protein secondary structure (PSS) prediction.

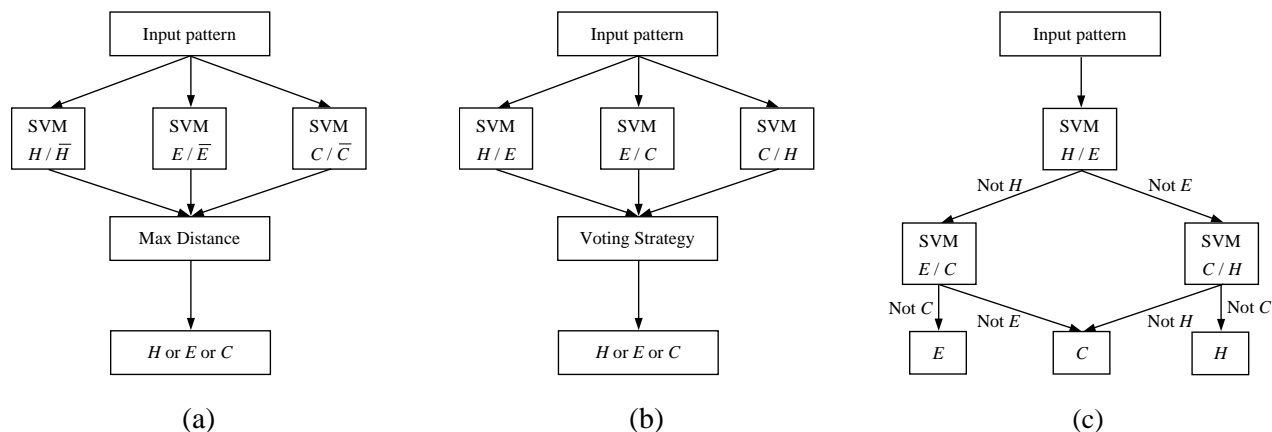


Figure 1: Illustration of techniques of PSS prediction using SVM approaches (a) one-against-all (OAA), (b) one-against-one (OAO), and (c) directed-acyclic-graph (DAG) methods.

2.1 One-Against-All (OAA) Method

A typical SVM model of one-against-all method [18] used for PSS prediction is illustrated in figure 1(a). Three binary SVM classifiers, H/\bar{H} , E/\bar{E} , C/\bar{C} are constructed, each predicting that the secondary structure at the local site i belongs to a particular secondary structure type or not. The input vectors, usually derived from a window of 7-15 amino acid residues, are transformed to a high dimensional space and compared to the support vectors via a kernel function for each classifier. The results are then linearly combined by using parameters α_i , $i = 1, 2, \dots, n$ that are found by solving a quadratic optimization problem. The process of training of each classifier k/\bar{k} , $k \in \Sigma_T$ is illustrated in the algorithm I:

Algorithm I: Classifier k/\bar{k}

Inputs: Training examples $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ and class labels $\{q_1^k, q_2^k, \dots, q_n^k\}$ where $q_i^k \in \{+1, -1\}$.

Maximize over α_i^k :

$$Q^k = \sum_{i=1}^n \alpha_i^k - (1/2) \sum_{i=1}^n \sum_{j=1}^n \alpha_i^k \alpha_j^k q_i^k q_j^k \mathcal{K}(\mathbf{v}_i, \mathbf{v}_j)$$

subject to

$$0 \leq \alpha_i^k \leq \gamma \text{ and } \sum_{i=1}^n \alpha_i^k q_i^k = 0$$

Outputs: Parameters α_i^k

The algorithm I, when the cost function Q^k is optimized, yields a classifier for k/\bar{k} with maximum margin of separation [20]. The summations of the maximizing function Q^k run over all training patterns. $\mathcal{K}(\mathbf{v}_i, \mathbf{v}_j) = \phi(\mathbf{v}_i)\phi(\mathbf{v}_j)$ denotes the kernel function where ϕ is a mapping used to convert input vectors \mathbf{v}_i into a high dimensional space, q_i^k encodes the secondary structure such that a binary value +1 if the secondary of the residue r_i is the secondary structure k or -1 otherwise, and γ is a positive constant used to decide the trade-off between training error and the margin. Once the parameters α_i are obtained from the above algorithm, the resulting discriminant function is known.

The resulting discriminant function of a new input vector \mathbf{v}_j of the above classifier is given by

$$D^k(\mathbf{v}_j) = \sum_{i=1}^n q_i^k \alpha_i^k \mathcal{K}(\mathbf{v}_i, \mathbf{v}_j) + b^k = \mathbf{w}^k \phi(\mathbf{v}_j) + b^k \quad (1)$$

where the bias b^k is chosen so that $q_i^k D^k(\mathbf{v}_i) = 1$ for any i with $0 < \alpha_i^k < \gamma$ and the weight vector $\mathbf{w}^k = \sum_{i=1}^n q_i^k \alpha_i^k \phi(\mathbf{v}_i)$. The secondary structural type t_j of the residue r_j is determined by the winner-take-all scheme, i.e. by taking the highest value of three discriminant function values.

$$t_j = \arg \max_{k \in \Sigma_T} D^k(\mathbf{v}_j) \quad (2)$$

2.2 One-Against-One (OAO) Method

This method constructs three binary classifiers, H/E , E/C , C/H , where each one is trained on data from two classes [10]. That is, for each pair k/t where $k, t \in \Sigma_T$ and $k \neq t$, a binary classifier is constructed which maps the examples with class k to +1 and the examples with class t to -1. The training process of each classifier k/t is the same as the training process of the one-against-all method where q_i^k encodes the secondary structure such that a binary value +1 if the secondary of the residue r_i is the secondary structure k or -1 if the secondary of the residue r_i is t . After training classifier k/t , the secondary structural type t_j of a new residue r_j is considered as k if the discriminant function $D^k(\mathbf{v}_j) \geq 0$ or t otherwise. Finally, a majority voting scheme for the signed output of the classifiers is employed to determine the class of the new object r_j : if classifier k/t says r_j is in the class k , the vote for the class k is added by one. Otherwise, the class t is increased by one. Then we predict r_j is in the class with the largest vote. The architecture of the one-against-one method for PSS prediction is in figure 1(b).

2.3 Directed Acyclic Graph (DAG) Method

In this method, the training phase is the same as the one-against-one method by solving three binary SVM classifiers, $H/E, E/C, C/H$ [14]. However, in testing phase, it uses a directed acyclic graph (DAG) that has three internal nodes and three leaves. Each node is a binary classifier k/t where $k, t \in \Sigma_T$ and $k \neq t$. Given a test residue r_j , starting at the root node, the binary decision function of the classifier H/E is evaluated. The node is then exited via the left edge, if the secondary structural type of the residue r_j is not a helix (H); or the right edge if its secondary structure is not a strand (E). The next node's decision function is then evaluated. Finally, the input r_j reaches to a leaf node that indicates the predicted secondary structure. The architecture of the DAG method for PSS prediction is in figure 1(c).

2.4 Vapnik and Weston (VW) Method

Vapnik and Weston have proposed an approach for multi-class problems by solving one single optimization problem [20, 21]. The idea is similar to the one-against-all approach. For PSS prediction, it constructs three two-class rules where the $k \in \Sigma_T$ th discriminant function $\mathbf{w}^k \phi(\mathbf{v}) + b^k$ separates training vectors of the class k from the other vectors. Basically, this method solves the following primal problem:

Minimize

$$\frac{1}{2} \sum_{k \in \Sigma_T} (\mathbf{w}^k \mathbf{w}^k) + \gamma \sum_{i=1}^n \sum_{k \neq k_i} \xi_i^k$$

subject to the constraints

$$\begin{aligned} \mathbf{w}^{k_i} \phi(\mathbf{v}_i) + b^{k_i} &\geq \mathbf{w}^k \phi(\mathbf{v}_i) + b^k + 2 - \xi_i^k \\ \xi_i^k &\geq 0 \end{aligned}$$

where k_i is the secondary structure of the residue r_i , ξ_i^k represents the distance from the correct side of the hyperplane (k_i, k) of the training vector \mathbf{v}_i , $i = 1, 2, \dots, n$, and $k \in \Sigma_T \setminus k_i$

Like binary SVM, minimization of this function can be done by solving the following convex quadratic programming (QP) problem.

$$\max_{\alpha} 2 \sum_{i=1}^n \sum_{k \in \Sigma_T} \alpha_i^k - \sum_{i=1}^n \sum_{j=1}^n \left(\frac{1}{2} c_j^{k_i} A_i A_j - \sum_{k \in \Sigma_T} \alpha_i^k \alpha_j^{k_i} + \frac{1}{2} \sum_{k \in \Sigma_T} \alpha_i^k \alpha_j^k \right) \mathcal{K}(\mathbf{v}_i, \mathbf{v}_j)$$

$$\text{such that } \sum_{i=1}^n \alpha_i^k = \sum_{i=1}^n c_i^k A_i, \quad 0 \leq \alpha_i^k \leq \gamma, \quad \text{and } \alpha_i^{k_i} = 0$$

$$\text{where } A_i = \sum_{k \in \Sigma_T} \alpha_i^k, \quad c_i^k = \begin{cases} 1 & \text{if } k_i = k \\ 0 & \text{if } k_i \neq k \end{cases}, \quad \text{and } \mathbf{w}^k = \sum_{i=1}^n (c_i^k A_i - \alpha_i^k) \phi(\mathbf{v}_i)$$

Once the parameters α_i^k are obtained from the above optimization, the resulting discriminant function of a new input vector \mathbf{v}_j is given by

$$D^k(\mathbf{v}_j) = \sum_{i=1}^n (c_i^k A_i - \alpha_i^k) \mathcal{K}(\mathbf{v}_i, \mathbf{v}_j) + b^k = \mathbf{w}^k \phi(\mathbf{v}_j) + b^k \quad (3)$$

The estimate of the secondary structural type at the site j is determined by the highest value of three discriminant functions

$$t_j = \arg \max_{k \in \Sigma_T} D^k(\mathbf{v}_j) \quad (4)$$

2.5 Crammer and Singer (CS) Method

Another approach for multi-class scheme has been proposed by Crammer and Singer [2]. For secondary structure prediction, this method constructs three discriminat functions but all are obtained by solving one single optimization problem. The formulation is as follows:

Minimize

$$\frac{1}{2} \sum_{k \in \Sigma_T} (\mathbf{w}^k \mathbf{w}^k) + \gamma \sum_{i=1}^n \xi_i$$

subject to the constraints

$$\mathbf{w}^{k_i} \phi(\mathbf{v}_i) - \mathbf{w}^k \phi(\mathbf{v}_i) \geq e_i^k - \xi_i$$

where k_i is the secondary structural type of the residue r_i , $i = 1, 2, \dots, n$, $k \in \Sigma_T$, and $e_i^k = 1 - c_i^k$, $c_i^k = \begin{cases} 1 & \text{if } k_i = k \\ 0 & \text{if } k_i \neq k \end{cases}$

The main difference from the method of Vapnik and Weston is that in this method only n slack variables ξ_i are used. In addition, the method does not have coefficients b^k , $k \in \Sigma_T$.

We find the minimization of the above function by solving the following QP problem:

$$\max_{\alpha} \sum_{i=1}^n \sum_{k \in \Sigma_T} \alpha_i^k e_i^k - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathcal{K}(\mathbf{v}_i, \mathbf{v}_j) \sum_{k \in \Sigma_T} (\gamma c_i^k - \alpha_i^k)(\gamma c_j^k - \alpha_j^k) \quad (5)$$

$$\text{such that } \alpha_i^k \geq 0, \quad \sum_{k \in \Sigma_T} \alpha_i^k = \gamma, \quad i = 1, 2, \dots, n$$

$$\text{where } \mathbf{w}^k = \sum_{i=1}^n (\gamma c_i^k - \alpha_i^k) \phi(\mathbf{v}_i)$$

To simplify this equation we denote $\beta_i^k = \gamma c_i^k - \alpha_i^k$. Eq. (5) becomes

$$\max_{\beta} - \sum_{i=1}^n \sum_{k \in \Sigma_T} \beta_i^k e_i^k - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathcal{K}(\mathbf{v}_i, \mathbf{v}_j) \sum_{k \in \Sigma_T} \beta_i^k \beta_j^k \quad (6)$$

$$\text{such that } \sum_{k \in \Sigma_T} \beta_i^k = 0, \quad i = 1, 2, \dots, n \text{ and } \beta_i^k \leq \begin{cases} 0 & \text{if } k_i \neq k \\ \gamma & \text{if } k_i = k \end{cases}$$

$$\text{where } \mathbf{w}^k = \sum_{i=1}^n \beta_i^k \phi(\mathbf{v}_i)$$

Once the parameters β_i^k are obtained from the optimization, the resulting discriminant function of a new input vector \mathbf{v}_j is given by

$$D^k(\mathbf{v}_j) = \sum_{i=1}^n \beta_i^k \mathcal{K}(\mathbf{v}_i, \mathbf{v}_j) = \mathbf{w}^k \phi(\mathbf{v}_j) \quad (7)$$

The secondary structural type at the site j is determined by

$$t_j = \arg \max_{k \in \Sigma_T} D^k(\mathbf{v}_j) \quad (8)$$

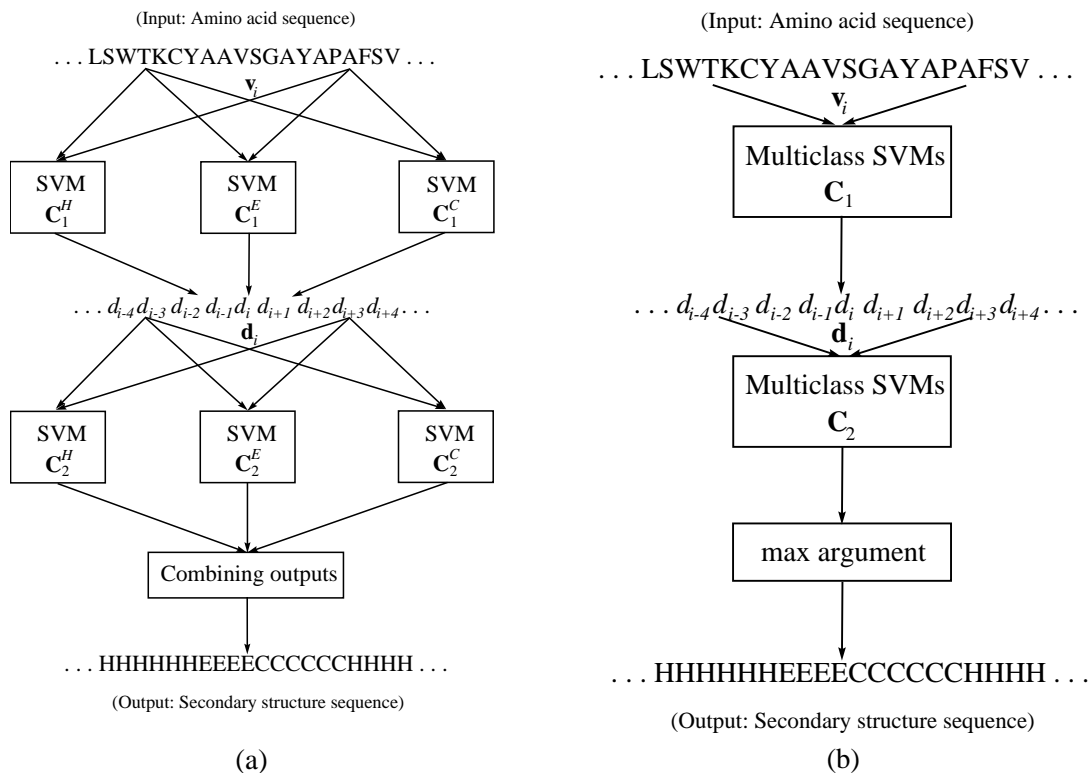


Figure 2: Illustration of two-stage SVM approaches for PSS using (a) binary classifiers and (b) multi-class techniques.

3 Two-Stage SVM Approaches

In this section, we present an architecture for protein classification by cascading two SVM classifiers.

In the above single-stage SVM approaches, the discriminant function values indicating belongingness of the input pattern to secondary structural elements were obtained. The technique, however, depends on the size of the input (window) and the errors may be introduced due to non-optimal values of parameters. We propose another SVM cascade to the first SVM to predict the secondary structure element using the secondary structure sequence given by the first-stage SVMs. The tenet of our approach is that the risk involved in the single-stage approaches could be minimized by incorporating the contextual information of the output secondary structure sequences, and hence the long range interactions of amino acid.

Consider a window of w_2 size around an element at the output of the first stage and the vector $\mathbf{d}_i = (d_{i-h_1}^k, d_{i-h_1+1}^k, \dots, d_i^k, \dots, d_{i+h_2}^k)$ where $w_2 = 3(h_1+h_2+1)$, $d_i^k = 1/(1+e^{-D^k(\mathbf{v}_i)})$, and $k \in \Sigma_T$ at site i . In two-stage SVM approaches for PSS prediction, the second stage combines the output of first stage SVMs to predict PSS. The figure 2 represents two schemes for PSS prediction: one combining two binary SVMs and one combining two multi-class SVMs. In binary SVMs, we use three classifiers, C_2^k , $k \in \Sigma_T$, each deciding whether or not a particular element belongs to a particular, i.e k , secondary structure. For this purpose, we use three SVMs associated with the three structuring elements and their outputs are combined to obtain the decision of the secondary structure. Our approach for binary SVMs is illustrated in figure 2(a). For OAA method, C^k is the classifier k/\bar{k} , $k \in \Sigma_T$ and the outputs of second stage SVMs are combined by the winner-take-all scheme. In OAO and DAG approaches, C^H, C^E, C^C are the classifiers $H/E, E/C, C/H$ and the outputs of second-stage SVMs are combined by the voting and DAG strategies, respectively. In multi-class SVM techniques, the input vectors \mathbf{d}_i are directly used for another multi-class SVM classifier at the second stage. The outputs of second stage

multi-class SVMs are combined by the winner-take-all scheme. Figure 2(b) illustrates our approach for multi-class SVMs.

Our approach supports that it is possible to improve the prediction accuracy by second SVM classifiers at output of existing secondary structure prediction scheme [15]. This is because the secondary structure at a particular position of the sequence depends on the structures of the rest of the sequence. This intrinsic relation cannot be captured by using only the single stage approaches alone. Therefore, another layer of classifiers which minimize the risk of the output of single stage methods improves the prediction accuracy. As shown, SVMs are optimal classifiers for the second stage because they minimize not only the empirical risk of known sequences but also the actual risk of unknown sequences.

Table 1: Comparison of performances of two-stage and single-stage SVM approaches in protein secondary structure prediction on RS126 dataset with multiple sequence alignments and the following reduction: H, G to (H); E, B to (E); the remainder to (C).

Method	Q_3 (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C	Sov (%)
<i>Single-Stage</i>								
One-against-all	70.4	69.7	54.1	79.3	0.59	0.51	0.50	59.0
One-against-one	70.1	67.6	54.5	79.8	0.59	0.50	0.49	58.3
DAG	70.1	67.5	54.2	80.0	0.59	0.50	0.49	58.3
Vapnik and Weston	70.5	70.4	55.7	78.2	0.61	0.51	0.49	57.1
Crammer and Singer	70.4	70.2	55.8	78.0	0.60	0.51	0.49	56.5
<i>Two-Stage</i>								
One-against-all	72.5	66.5	61.2	78.5	0.62	0.55	0.52	63.9
One-against-one	72.1	66.5	57.5	81.2	0.60	0.55	0.50	65.4
DAG	72.1	66.8	57.4	80.9	0.59	0.55	0.51	65.5
Vapnik and Weston	72.8	66.1	57.8	81.9	0.63	0.55	0.50	67.0
Crammer and Singer	72.7	66.8	57.9	81.0	0.62	0.55	0.50	66.8

4 Experiments and Results

4.1 Datasets

The set 126 nonhomologous globular protein chains used in the experiment of Rost and Sander (Rost et al. 1993), referred to as the RS126 set, was used to evaluate the accuracy of the classifiers. The dataset contained 23349 residues with 32% α -helix, 23% β -strand, and 45% coil. Many current generation secondary structure prediction methods have been developed and tested on this dataset. The RS126 set is available at <http://www.compbio.dundee.ac.uk/~www-jpred/data/>. The single-stage approaches and second-stage approaches were implemented, with multiple sequence alignments, and tested on the dataset, using a sevenfold cross validation technique to estimate the prediction accuracy. With sevenfold cross validation approximately one-seventh of the database was left out while training and, after training, the left one-seventh of the dataset was used for testing. In order to avoid the selection of extremely biased partitions, the RS126 set was divided into seven subsets with

Table 2: Comparison of performances of two-stage and single-stage multi-class SVM approach with position specific scoring matrices generated by PSI-BLAST on PSIPRED dataset and the following reduction: H to (H); E to (E); the remainder to (C).

Method	Q_3 (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C	Sov (%)
<i>Single-stage</i> Vapnik and Weston	77.7	74.3	55.2	88.7	0.72	0.56	0.58	70.5
<i>Two-stage</i> Vapnik and Weston	79.5	76.7	60.2	87.4	0.74	0.60	0.61	76.3

each subset having similar size and content of each type of secondary structure.

A two-stage SVMs has been used to predict PSS based on the position specific scoring matrices generated by PSI-BLAST. By using the PSI-BLAST profiles directly, the PSIPRED method [9] achieved the highest published score for any previous methods. The testing set of 187 protein chains from PSIPRED method, which is available at <ftp://bioinf.cs.ucl.ac.uk/pub/psipred/old/data/>, was used to evaluate the accuracy of two-stage SVMs.

4.2 Results

For SVM classifiers at the first stage, a window size of 11 amino acid residues ($h_1 = h_2 = 5$) was used as input for optimal result in the [7, 15] range. At the second stage, the window size of width 21 ($h_1 = 2$ and $h_2 = 4$) in the [9, 24] range gave the optimal accuracy of all second-stage SVM techniques. The kernel selected here was the radial basis function $\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\sigma \|\mathbf{x} - \mathbf{y}\|^2}$ with the parameters: $\sigma = 0.25$, $\gamma = 2.0$ on RS126 and $\sigma = 0.05$, $\gamma = 2.0$ on PSIPRED set at the first stage, and $\sigma = 0.001$, $\gamma = 1.0$ on RS126 and $\sigma = 0.01$, $\gamma = 2.5$ on PSIPRED set at the second stage, determined empirically for optimal performance. The use of Gaussian kernel showed the best performance even though the dimension of feature space is infinite [19]. The one-against-all, one-against-one, and DAG methods were implemented using sequential minimization algorithm [13] which is simple to implement without needing storage for matrices or to invoke an iterative numerical routine for each sub-problem. We used BSVM library [7], which leads to faster convergences for large optimization problem, to implement two multi-class techniques.

We have used several measures to evaluate the prediction accuracy. The Q_3 accuracy indicates the percentage of correctly predicted residues of three states of secondary structure [3]. The Q_H, Q_E, Q_C accuracies represent the percentage of correctly predicted residues of each type of secondary structure [3]. Matthew's correlation coefficients (ρ_H, ρ_E, ρ_C) provide the success of predicting residues for each type of secondary structure [11]. Segment overlap measure (Sov) gives accuracy by counting predicted and observed segments, and measuring their overlap [3].

Table 1 shows the performance of the different secondary structure predictors using two-level SVMs on the RS126 set with multiple sequence alignments. At the first stage, the multi-class techniques of Vapnik and Weston gave the best result for PSS prediction which achieved 70.5% of Q_3 accuracy while the accuracy of one-against-all, one-against-one, DAG, and Crammer and Singer methods were 70.4%, 70.1%, 70.1%, and 70.4% respectively. The best algorithm was found to be the cascade of two SVMs

using the multi-class techniques of Vapnik and Weston, which achieved 72.8% of Q_3 accuracy while the prediction accuracy made by Rost and Sander's PHD method was only 70.8%. Table 2 shows the performance of two-stage SVMs with the PSIPRED dataset based on PSI-BLAST profile. The best prediction was achieved to be combination of the Vapnik and Weston's multi-class SVM and SVM: 79.5% of Q_3 accuracy while the accuracy of PSIPRED method was previously reported to be 78.3%. By using different binary SVM classifiers and multi-class SVM techniques to enhance the prediction of the SVM secondary structure schemes at the first stage, the new prediction schemes achieved 2% of improvement in the Q_3 accuracy and 5-11% of improvement in the Sov accuracy.

5 Discussion and Conclusion

We have compared most multi-class SVMs currently reported in the literature for PSS problem, which are three methods based on binary classifications: one-against-all, one-against-one, and directed acyclic graph, and two approaches for multi-class problem by solving one single optimization problem. We found that the multi-class SVMs proposed by Vapnik and Weston are more suitable for protein secondary structure prediction than the other methods, including binary SVMs, because their capacity to solve the optimization problem in one step.

Furthermore, we introduced a general framework for two-stage approaches by using SVMs to predict PSS from the output from earlier single-stage SVM techniques. Our experiments demonstrated that it is feasible to extend current single-stage approaches with a second-stage to improve the accuracy of prediction because secondary structure at a particular position of a sequence depends not only on the amino acid residue at a particular location but also on the structural formations of the rest of the sequence. This intrinsic relation cannot be captured by using only single-stage approaches alone. Therefore, another layer of classifiers, which predicts the output of single-stage methods, improves the accuracy of prediction. As seen in the experiments, SVMs were optimal classifiers for the second-stage because they minimized not only the empirical risk of known sequences but also the actual risk of unknown sequences. Additionally, two stages were sufficient to find an optimal classifier for PSS prediction as SVMs minimized the generalization error of the output of single stage by solving the optimization problems at second stage.

References

- [1] Clote, P. and Backofen, R., *Computational Molecular Biology*, Wiley and Sons, Ltd., Chichester, 2000.
- [2] Crammer, K. and Singer, Y., On the learnability and design of output codes for multi-class problems, *Computational Learning Theory*, 35–46, 2000.
- [3] Cuff, J.A. and Barton, G.J., Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, *Proteins*, 4:508–519, 1999.
- [4] Garnier, J., Osguthorpe, D.J., and Robson, B., Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *Journal of Molecular Biology*, 120:97–120, 1978.
- [5] Garnier, J., Gibrat, J.F., and Robson, B., GOR method for predicting protein secondary structure from amino acid sequence, *Methods Enzymol*, 266:541–553, 1996.
- [6] Gibrat, J.F., Garnier, J., and Robson, B., Further developments of protein secondary structure prediction using information theory, *Journal of Molecular Biology*, 198:425–443, 1987.

- [7] Hsu, C.W. and Lin. C.J., A comparison on methods for multi-class support vector machines, *IEEE Transactions on Neural Networks*, 13:415–425, 2002.
- [8] Hua, S. and Sun, Z., A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, *Journal of Molecular Biology*, 308:397–407, 2001.
- [9] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices, *Journal of Molecular Biology*, 292:195–202, 1999.
- [10] Kreßel, U., Pairwise classification and support vector machines, *In Advances in Kernel Methods-Support Vector Learning*, Cambridge, MA: MIT Press, 255–268, 1999.
- [11] Matthews, B., Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta*, 405:442–451, 1975.
- [12] Mount, D.W., *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2001.
- [13] Platt, J.C., Using sparseness and analytic QP to speed training of support vector machines, *Advances in Neural Information Processing Systems 11*, Cambridge, MA:MIT Press, 1999.
- [14] Platt, J.C., Cristianini, N., and Shawe-Taylor, J., Large margin DAG's for multiclass classification, *Advances in Neural Information Processing Systems 12*, Cambridge, MA: MIT Press, 12:547–553, 2000.
- [15] Rost, B. and Sander, C., Prediction of protein secondary structure at better than 70% accuracy, *Journal of Molecular Biology*, 232:584–599, 1993.
- [16] Salamov, A.A. and Solovyev, V.V., Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments, *Journal of Molecular Biology*, 247:11–15, 1995.
- [17] Salamov, A.A. and Solovyev, V.V., Protein secondary structure prediction using local alignments, *Journal of Molecular Biology*, 268:31–36, 1997.
- [18] Scholköpfung, B., Burges, C., and Vapnik, V., Extracting support data for a given task, *Proc. First International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1995.
- [19] Scholköpfung, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V., Comparing support vector machines with gaussian kernels to radial basis function classifiers, *IEEE Trans. Sign. Processing*, 45:2758–2765, 1997.
- [20] Vapnik, V., *Statistical Learning Theory*, Wiley and Sons, Inc., New York, 1998.
- [21] Weston, J. and Watkins, C., Multi-class support vector machines, In Verleysen, M., editor, *Proceedings of ESANN99, Brussels*, D. Facto Press, 1999.