Prediction of Protein Secondary Structure Using LIBSVM and SSVM with different Kernel Function

***[1] Rajnish K. Vashishtha, [2] Abhay banker, [3] Neetu Verma, [4] C. K. Verma , [5] Manoj Jha**
[1, 2, 3, 4, 5] Department of Mathematics, Bioinformatics and Computer Applications
Maulana Azad National Institute of Technology Bhopal, India
** Email: rajnishmanit@gmail.com, Contact: +919584669064*

Received: 24th Aug 2016, Accepted: 26th Aug 2016, Published: 1st Sep 2016

Abstract
Bioinformatics and computational biology is the emerging field of research in current era. Proteomics is solitary of the prime areas of research in bioinformatics. The general function of proteins is mostly dictated by structure. Prediction of secondary structure plays a key role in its further prediction of tertiary structure. Several techniques have been proposed by the researchers for protein secondary structure prediction. Since the data is inherently complex and these approaches endure the problem of accuracy. There are lots of methods based on machine learning procedure, such as neural networks (NN), clustering, decision tree and LIBSVM etc. Support Vector Machine has exposed physically powerful generalization capability of prediction of protein structure. In this paper, the important point on protein secondary structure by binary classification using LIBSVM and SSVM with the diverse kernel role has been detailed. The proposed method gives a better accuracy than existing technique for the prediction of protein structure. This method can be performed in MATLAB software. Two datasets are used for this problem optimization.

Keywords:
Protein Secondary Structure Prediction, Zika Virus Datasets, Support Vector Machine, Smoothing and Kernel Function

Introduction

Bioinformatics and data mining is the highest rising area of research. Proteomics is one of the richest area in Bioinformatics because protein plays significant task in a biological process, understanding the function, discovery of novel medicine and products of industrial and medical application. Many researchers attempt to determine the construction of proteins using nuclear magnetic resonance and X-ray crystallography. Both methods are time consuming & expensive [1]. Prior information on class of protein structure will significantly pick up the quality and piece of secondary structure prediction of protein from amino acid succession by dipping the search space of the structure prediction process. The idea of protein structural class was first introduced by Chortia and Levitt [2]. In general protein structure can be classified into α, β, α/β, α+β. A protein is formed by the specific sequence of amino acids. This linear string is called the primary structure. The protein secondary structure help to resolve of tertiary structure by fold recognition method. Main essential parts of secondary structure are coil or turn, α- helix, β-sheets. Third level in structure of protein is multimeric and monomeric protein molecules. It's very difficult to predict. In this paper we are center on α- helix, β-sheets for binary categorization of protein secondary structure. Our objective is to travel around the problem of Protein structure forecast using machine learning techniques.

The key idea of machine learning is intend mechanism to learn like a human, study from knowledge and discover information from dataset. Various machine learning technique are used to handle prediction of protein structure problem in bioinformatics [3]. The study attempted to concern support vector machine and smooth support vector machine using a linear, quadratic, polynomial and RBF kernel functions for binary classification of protein secondary structure. These methods are applied on dataset. These dataset sequences are derived from NCBI & use Chou fasman algorithms. As a result smooth support vector machine reached high accuracy, this shows that the structure class of a protein is considerably correlated with its amino acid composition. The smooth support vector machine can become a helpful for forecast the structural classes of protein.

Materials and Methods

2.1. Data set

In this study, we have used NS5 and Polyprotein of Zika virus dataset. The all dataset are prepared using from NCBI and Chou fasman algorithms. The first dataset NS5 Zika virus contains 274 domains, of which 197 are all alpha-helix and 77 are all beta-sheet. The second dataset polyprotein Zika virus contains 1175 domains, 585 all alpha-helix and 590 beta-sheets. All dataset are training and testing from LIBSVM and SSVM in MATLAB environment.

2.2. Nonlinear LIBSVM

The LIBSVM has significantly provided better performance than other traditional machine learning technique like as neural networks (NNs), clustering etc [4]. LIBSVM is the kind of learning machine depend on statistical learning theory and investigate

data used for regression and classification [5]. Basically LIBSVM is applied model classification in different stages as follows: first mapping the input vectors into feature space, either linear or non-linear, which is pertinent with the abundance of kernel functions, then optimized Linear kernel function is making a hyper plane and break up two classes and nonlinear LIBSVM is extended to multiple class [6].LIBSVM training explanation is a global optimized solution. LIBSVM is used in a lot of research area like as drug design, image recognition and classification [7].

In this proposed work we have applied LIBSVM for forecasting the structural classes of protein [8]. In MATLAB software used LIBSVM for the problem optimization of protein structure prediction [9].

We are given a training dataset of instance pairs (x_i, y_i) i=1.....m where $x_i \in R^n$ and $y_i \in (1,0)$
LIBSVM involve the explanation of subsequent problems.

$\min_{w,b,\gamma} \frac{1}{2} \| w \|^2$…… (1)

Subject to

$y_i(w^T(x_i) + b) \geq 1$ …… (2)

In gradient constraint

$g_i(w) = -y_i(w^T(x_i) + b) + 1 \leq 0$……... (3)

In dual complementarily condition, that is $g_i(w) = 0$ and converted inner products $(x^T x)$ to(x_i^T, x_j) than, apply the kernel function.

$l(w, b, \alpha) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{m} \alpha_i y_i (w^T(x_i) - b) - 1$ …… …….. (4)

Where α_i is Lagrange multipliers, we have used a lagrangian method for finding he dual shape of this problem. than require to first reduce $l(w, b, \alpha)$ with respect to w and b (for fixedα_i), derivatives of l with respect to w and b to zero[15].
than

$w = \sum_{i=1}^{m} \alpha_i y_i \emptyset(x_i)$……..…. (5)

Substituting $w = \sum_{i=1}^{m} \alpha_i y_i \emptyset(x_i)$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$ into equation (4)
Now obtain the decision function

$f(x) = sign(w^T x - b)$ ……..…..…..…. (6)

$f(x) = sign(\sum_{i=1}^{m} \alpha_i y_i x^T x - b)$……..…..…… (7)

$f(x) = sign(\sum_{i=1}^{m} \alpha_i y_i (x_i^T x_j) - b)$…… (8)

Equation No. (8) Called decision function [15].

2.3 **Smooth support vector machine**

Smoothing process has been extensively worn for solving significant mathematical programming problems [13, 14].
We consider a problem is classified in training set $T = (x_i, y_i)$ where $x_i \in R^n$ and label set $L = (y_i)$ where $y_i \in (1, 0)$

$\min_{(w,\gamma,\xi)\epsilon R^{n+1+m}} \frac{1}{2} w'w + ve'y$

S.t. $D(Aw - e\gamma) + y \geq e$

$y \geq 0$

Where:
V= positive vector
y= slack variable
e= column vector
w= the normal weight of the bounding planes

$x'w - \gamma = +1$……..…… (9)

$x'w - \gamma = -1$ ……..... (10) γ

Determine their position compared to the origin. The linear separating surface is the plane

$x'w = \gamma$ …….. (11)

If classes are linearly inseparable, the bounding plans as follows-

$x'w - \gamma + y_i \geq +1$, for $x' = A_i$ and $D_{ii} = +1$….. (12) $x'w - \gamma + y_i \leq -1$, for $x' = A_i$ and $D_{ii} = -1$….. (13)

These constraints (12, 13) equation can be written as a matrix form as $D(Aw - e\gamma) + y \geq e$…… (14)

In SSVM approach the customized LIBSVM problem is given as

$\min_{(w,\gamma,\xi)\epsilon R^{n+1+m}} \frac{1}{2}(w'w + \gamma) + \frac{v}{2} y'y$

Subject to

$D(Aw - e\gamma) + y \geq e$…… (15)

$y \geq e$

The restraint in eq. (15) can be writing by

$y = (e - D(Aw - e\gamma))_+$…..…… (16)

Thus, we can substitute y in constraint (15) by (16) and exchange the SSVM problem (15) into an equivalent LIBSVM which is an unrestrained optimization problem as

$\min_{(w,\gamma,)} \frac{1}{2}(w'w + \gamma) + \frac{v}{2} \| (e - D(Aw - e\gamma))_+ \|_2^2$… (17)

Plus function $(x)_+$ is defined as

$(x)_+ = \max(0, x_i)$ …….. (18)

Where i=1, 2, 3……..n

The objective function in (18) is undifferentiable and unsmooth. Therefore it can't be resolve using conventional optimization method, because it always required that the objective functions gradient and hessian matrix.
Lee et al [10] applies the smoothing procedure and substitute $(x)_+$ by the integral of the sigmoid function. $p(x, \propto) = x + \frac{1}{\propto} \log(1 + \varepsilon^{-\alpha x}), \alpha > 0$…..…. (19)

This p function with a smoothing restriction α is used here to substitute the plus+ function of (17) to obtain a smooth support vector machine [11].

$$\min_{(w,\gamma,\xi)\in R^{n+1}} \quad \frac{1}{2}(w'w+\gamma)+\frac{v}{2}\parallel p(e-D(Aw-e\gamma),\alpha)\parallel_2^2 \ldots\ldots (20)$$

This solution of problem (15) is getting by solving problem (20) with α approaching infinity. The problem (20) can be resolved using a Newton-Armijo algorithm.

For nonlinear unseparable problem require a choosing kernel function K to reflect the input space into another space This representation was derive from generalized support vector machines[12]. So the problem (20) can be approximated as subsequent:

$$\min_{(u,y,\gamma)} \frac{1}{2}(u'u+\gamma^2)+\frac{v}{2}y'y$$

Subject to $D((k(A,A')Du-e\gamma)+y\geq e\ldots\ldots\ldots\ldots (21)$

$$y\geq 0$$

Same as previous, it is acquiring the SSVM for indivisible problem:

$$\min_{(u,\gamma,)} \quad \frac{1}{2}(u'u+\gamma^2)+\frac{v}{2}\parallel p(e-D((k(A,A')Du-e\gamma),\alpha)\parallel_2^2 .. (22)$$

Where, K (A, A′) is kernel map. We can furthermore be appropriate the Newton-Armijo algorithm straight to solve equation (22).

2.4. Kernel substitution

LIBSVM and SSVM are performing a non-linear mapping of input vectors, where the mapping is resolute by the kernel function. In section 2.2 and 2.3 define a non-linear decision function in the input space. In decision function substitute two diverse kernels given below.

Linear kernel -

$$k(x_i,x_j)=\emptyset(x_i).\emptyset(x_j)$$

RBF kernel -

$$k(x_i,x_j)=\exp(-\gamma\parallel x_i-x_j\parallel)$$

Results and Discussion:

3.1. Success rate of LIBSVM

In this proposed work, the linear, quadratic, polynomial and RBF kernels are used in LIBSVM approach. For the data validation, 10-fold cross validation tested in two different datasets from Zika virus. First dataset NS 5 contain 274 residues and polyprotein dataset contain 1175 residues. In NS 5 dataset, linear, quadratic, polynomial and RBF kernel function gives 55%, 67%, 75% and 72% mean accuracy respectively.

In polyprotein dataset, linear, quadratic, polynomial and RBF kernel function gives 52% 60%, 72% and 70 % mean accuracy respectively.

[Table-1] Success Rate of LIBSVM

S. No.	LIBSVM with diff. kernel function	Mean Accuracy in %	
		NS5 dataset	Poly protein dataset
1	Linear	55	52
2	Quadratic	67	60
3	RBF	72	70
4	Polynomial	75	72

Figure 1 LIBSVM linear kernel function graph of polyprotein datasets

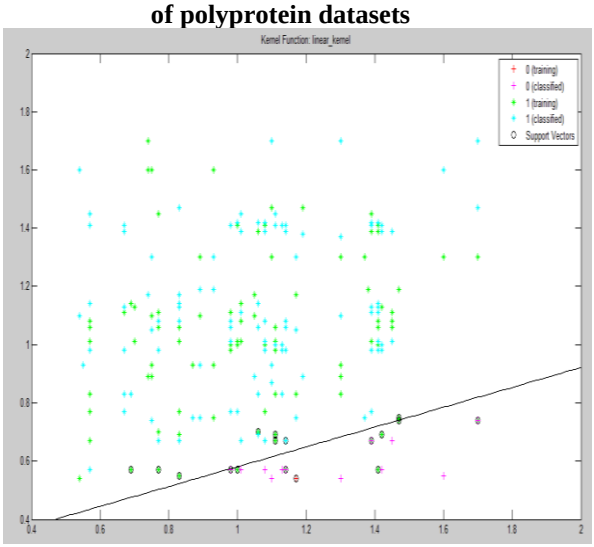


Figure 2 LIBSVM Quadratic kernel function graph of polyprotein dataset

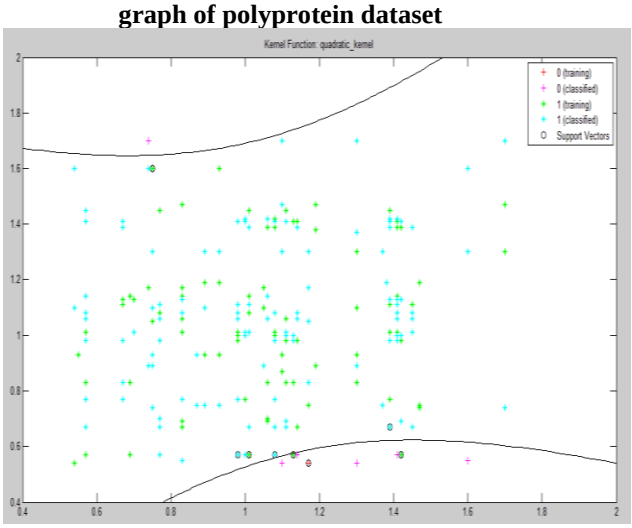


Figure 3 LIBSVM RBF kernel function graph of ploy protien dataset

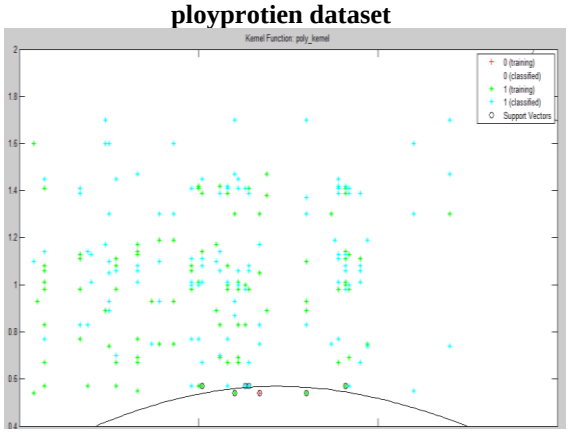
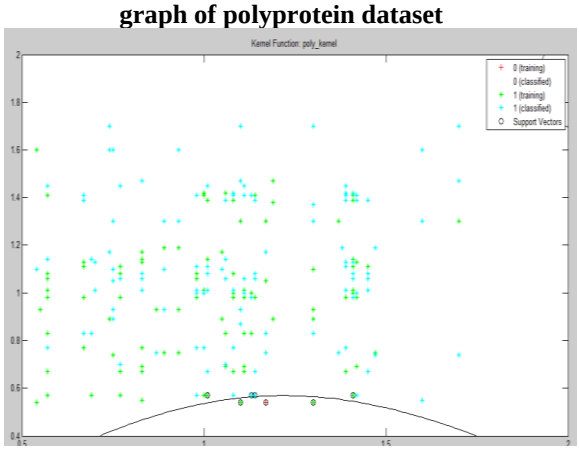


Figure 4 LIBSVM Polynomial kernel function graph of polyprotein dataset



3.2. Success rate of SSVM

The SSVM with linear and RBF kernel are two test often used 10-fold cross validation process. Among these two, the RBF test as the most effective. In this paper, the linear and RBF kernel with SSVM method was tested two different datasets that contain 274 and 1175 domains. Using linear kernel accuracy is 71% and 77% and using RBF kernel accuracy is 60% and 78%.

[Table-2] Success rate of SSVM

S. No.	SSVM with diverse kernel	Mean Accuracy in %	
		NS5 dataset	Poly protein dataset
1	Linear	71	59
2	RBF	77	78

3.2. Comparison of Success rate LIBSVM and SSVM

[Table-3] comparison of accuracy

S. No.	Ker-Nel	Mean Accuracy in %			
		LIBSVM dataset		SSVM datasets	
		NS5	Poly protein	NS5	Poly protein
1	Linear	55%	52%	71	60
2	RBF	72%	70%	77	78

Discussion

In this paper LIBSVM and SSVM are used for the classification. Linear, quadratic, polynomial & RBF, kernel functions are used for the classification accuracy of the data. In LIBSVM and SSVM , 10-fold cross validation technique is used for data validation. Smooth support vector machine mean accuracy is better than support vector machine mean accuracy in diverse kernel function. Smooth support vector machine (SSVM) provides a better accuracy for Zika virus data set. SSVM is helpful for protein structure prediction and classification**.**

Conclusion:

This research work is based on the forecasting of protein structure, and is useful for drug designing, drug discovery and data testing. All the prediction results are much closer to the different kernel functions. Also the error calculated to achieve SSVM is lesser than the error values obtained in LIBSVM approach. The SSVM approach takes less number of iterations to reach up to desired value and SSVM approach is significantly faster and provides better accuracy than LIBSVM approach. The SSVM can work with extensive range of data where input values are given with respective reference output values. Unlike LIBSVM, SSVM do not need a big training of data set. A small dataset is sufficient to train in SSVM and it becomes ready for prediction.

References

1. Chothia, Cyrus, and L. Michael. "Structural patterns in globular proteins."Nature 261 (1976): 552-558.

2. Shoyaib, Mohammad, et al. "Protein secondary structure prediction with high accuracy using support vector machine." Computer and information technology, 2007. iccit 2007. 10th international conference on. IEEE, 2007.

3. Cai, Yu-Dong, et al. "Prediction of protein structural classes by support vector machines." Computers & chemistry 26.3 (2002): 293-296.

4. He, Jieyue, et al. "Rule generation for protein secondary structure prediction with support vector machines and decision tree." IEEE Transactions on nanobioscience 5.1 (2006): 46-53.

5. Cai, Yu-Dong, et al. "Prediction of protein structural classes by support vector machines." Computers & chemistry 26.3 (2002): 293-296.

6. Mangasarian, Olvi L., and David R. Musicant. "Successive overrelaxation for support vector machines." IEEE Transactions on Neural Networks 10.5 (1999): 1032-1037.

7. Collobert, Ronan, et al. "Large scale transductive LIBSVMs." Journal of Machine Learning Research 7.Aug (2006): 1687-1712.

8. Muhamud, Ahmed I., M. B. Abdelhalim, and Mai S. Mabrouk. "Extraction of prediction rules: Protein secondary structure prediction." Computer Engineering Conference (ICENCO), 2014 10th International. IEEE, 2014.

9. Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1-16.

10. Lee, Yuh-Jye, and Olvi L. Mangasarian. "SLIBSVM: A smooth support vector machine for classification." Computational optimization and Applications20.1 (2001): 5-22.

11. Jie, Zhang, Fan Xuhui, and Ban Dengke. "Smooth support vector machine based on circular tangent function." The Journal of China Universities of Posts and Telecommunications 23.1 (2016): 68-96.

12. Chen, Chunhui, and Olvi L. Mangasarian. "A class of smoothing functions for nonlinear and mixedcomplementarity problems." Computational Optimization and Applications 5.2 (1996): 97-138.

13. Liang, Jinjin, and De Wu. "A new smooth support vector machine."International Conference on Artificial Intelligence and Computational Intelligence. Springer Berlin Heidelberg, 2010.

14. Yuan, Yubo, and Chunzhong Li. "A new smooth support vector machine."International Conference on Computational and Information Science. Springer Berlin Heidelberg, 2005.

15. Soman, K. P., R. Loganathan, and V. Ajay. Machine learning with SVM and other kernel methods. PHI Learning Pvt. Ltd., 2009.

