Fuzzy SVM with a Novel Membership Function for Prediction of Protein-Protein Interaction Sites in *Homo sapiens*

Brijesh Kumar Sriwastava[1], Subhadip Basu[2,*], and Ujjwal Maulik[2,*]

[1] Department of Computer Science and Engineering,
Government College of Engineering and Leather Technology, Kolkata-700098, India
{sriwastavabrijesh@yahoo.co.in}
[2] Department of Computer Science and Engineering, Jadavpur University,
Kolkata – 700032, India
{subhadip,umaulik}@cse.jdvu.ac.in

Abstract. Predicting residues that participate in protein–protein interactions (PPI) helps to identify the amino acids located at the interface. In this work, experimentally verified 3-D structures of protein complexes are used for building the training model and subsequent prediction protein interactions from sequence information. Fuzzy SVM (F-SVM), which is developed on top of the classical SVM, is an effective method to solve this problem and we demonstrate that the performance of the SVM can further be improved with the use of a custom-designed fuzzy membership function. We evaluate the performances of both SVM and F-SVM on the PPI database of the *Homo sapiens* organism and evaluate the statistical significance of F-SVM over classical SVM. To predict interaction sites in protein complexes, local composition of amino acids together with their physico-chemical characteristics are used. The F-SVM based residues prediction method exploits the membership function for each pair sequence fragment and in all cases F-SVM improves the performances obtained by the corresponding SVM classifiers. The F-SVM performance on the test samples is measured by area under ROC curve (AUC) as 80.16% which is around 1.55% higher than the classical SVM classifier.

Keywords: Protein–protein interaction, Support vector machine, Fuzzy SVM.

1 Introduction

Protein-protein interactions (PPI) are at the core of the entire interaction system of any living cell, making them the central hubs or major mediators for virtually every bio-chemical process. Two major types of complexes are observed, namely homodimers and heterodimers, where homodimers mostly form permanent and highly optimized complexes, generally by aligning hydrophobic interfaces. In contrast, in the case of hetero-complexes, hydrophobicity is indistinguishable from the rest of the surface [1-3]. Jones and Thornton [4] suggested importance of differentiating between

* Corresponding author.

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 668–673, 2013.

aforementioned types of complexes, when analysing their intermolecular interfaces. Summarizing, significant research was done in the area of protein-protein interactions, yet the problem of interaction sites prediction is still not fully understood.

Considering the inherent complexity of the problem, we have chosen to use SVM as the underlying classifier. However, the classical SVM algorithm is found to be inadequate to address the natural ambiguity in many datasets like the one used here for PPI site prediction. Therefore, we have used Fuzzy SVM (F-SVM) with a novel membership function to design the binary classification system for each pair sequence fragment to determine their interaction status.

Traditionally logic regression and neural networks technologies were used to deal with these noisy and ambiguous data sets. F-SVM, another alternative to work with noisy datasets, was first proposed in [5], where each sample is given a fuzzy membership that denotes the strength of belongingness of one data point towards one class.

Classical SVM implicitly uses the kernel function which often maps all training data from input space into a higher dimensional feature space. Fuzzy support vector machines (F-SVMs), Lin and Wang also worked in the same way, except that a membership value is associated with each training vector. Here, the membership value is multiplied into the penalty term to provide variable weighting [5].

In this paper we demonstrate that the performance of SVM algorithm can further be improved with the use of a custom-designed fuzzy membership function, for the PPI prediction problem. In this regard, we first discuss briefly the SVM classifier and then we highlight the design of the F-SVM classifier with the novel membership function for each pair sequence fragment to encode their interaction strength. Finally, we evaluate the performances of both SVM and F-SVM on the PPI databases of Homo sapiens.

2 Methods

2.1 Fuzzy Support Vector Machine Classifier

Support Vector Machine is an important machine learning technique proposed by Vapnik and co-workers [6]. In classical SVM, each sample is treated equally, *i.e.*, each input point is assigned to either one of the two classes. However, in some problems, some input points, such as the outliers, may not be exactly assigned to one of these two classes. In this context each point does not have the same meaning to the decision surface. In order to solve this problem, fuzzy membership of each input point may be introduced in such a way that different input points can make different contribution to the construction of decision surface. Suppose the training samples with associated fuzzy membership are $(x_1, y_1, s_1), (x_2, y_2, s_2), \ldots\ldots, (x_l, y_l, s_l)$, where each $x_i \epsilon R^N$ is a training sample, $y_i \epsilon \{+1, -1\}$ represents their class label and s_i is the fuzzy membership of point x_i which satisfies the condition $\sigma \leq s_i \leq 1$, $i = \{1,2,,\ldots.,l\}$ and $\sigma > 0$.

The fuzzy membership s_i is the attitude of the corresponding point x_i to belong one class and the parameter ξ_i is a measure of error in the SVM, the term $s_i\xi_i$ is a measure of error with different weighting to belong to that class. The problem of finding the optimal hyperplane can then be formulated as:

$$\begin{cases} \min & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} s_i \xi_i \\ y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0 & \forall \ i = 1,2,\ldots,l \end{cases} \quad (1)$$

where, C is a constant. It is important to note that if s_i is small then $s_i\xi_i$ also becomes small and effect of parameter ξ_i in (1) reduced. Therefore, corresponding point x_i have lesser impact on decision solution.

In F-SVM, greatest lower bound of α_i is zero which is same as in classical SVM. However the lowest upper bound for α_i is s_iC, which is not constant as in classical SVM. So feasible region of α_i dynamically depends on fuzzy membership value (s_i) of point x_i belonging to that class.

The training set $T = \{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_l, y_l)\}$, where $x_i \epsilon R^N$ belongs to one of the class $y_i \epsilon(+1, -1)$ for $i = 1,2, \ldots\ldots, l$. Subsequently a 2-dimensional matrix R is computed for distance of each vector to other vectors of set

$$R = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1l} \\ d_{21} & d_{22} & \cdots & d_{2l} \\ \vdots & \vdots & \vdots & \vdots \\ d_{l1} & d_{l2} & \cdots & d_{ll} \end{bmatrix} \tag{2}$$

From the matrix we find the maximum distance as $d = \max_{i,j=1,2,\ldots\ldots,j} d_{ij}$. Then average distance of all sample points is calculated as:

$$D = \frac{1}{C_l^2}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l} d_{ij} = \frac{2}{l(l-1)}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l} d_{ij} \tag{3}$$

Let the density of Sample Point be ρ_i and higher value of ρ_i signifies that more points are flanking to it and the corresponding to this points, the SVM have major role in classification, whereas the lower value of ρ_i means there are sparse points near to x_i and hence the role of SVM in classification is lesser. Then, we multiply the posterior probability with density ratio $\frac{\rho_i}{\rho}$ to reflect the possibility of the sample belongs to the class that constitutes and based on the posterior probability weighting membership function [7]. We need to understand the sample distribution position and can adjust samples in the role of classification which are near to the hyperplane. This ratio may be greater than 1 or less than 1 but when it less than 1, it signifies that point is outlier or noise point. Then fuzzy membership is defined as follows:

$$\mu(x_i) = P(\omega_j|x_i).\frac{\rho_i}{\rho} \quad , i = 1,2, \ldots\ldots, l \ \ and \ j = 1,2 \tag{4}$$

2.2 Fuzzy Membership Evaluation

In practice, we determine the membership function based on domain experience. We are working with 21 size window fragment (win_size) [8, 9] and we are considering a fragment to be positive if there are at least 2 interactions. The choice of two interactions is made to avoid possible noise, in the form of isolated interaction residue, in the positive dataset. Now maximum number of interactions in feature vector of size 21 is $21 \times 21 = 441$. But, every feature vector is not of equal strength with respect to number of interaction. So, it is better that each feature vector will not give equal contribution for training i.e. the feature vector having higher interaction strength is likely to be have more impact on training rather than lower strength feature vector. This is just due to the fact that, if there are more interactions in a fragment then more positive features are trained and gives better decision capability of choosing positive and negative class when the test would performed. This idea is used to give fuzzy membership strength to the each feature vector which is as below:

$$fs_i = \frac{num\_it}{(win\_size \times win\_size)} \qquad , i = 1,2, \dots \dots, l \tag{5}$$

where fs_i is feature strength of i^{th} vector, num_it is number of interaction in i^{th} feature vector and win_size is window size. Now, each feature vector has its own strength and we have taken average of all such vector and that is used to give weight among all vectors. This is separately done for $+ve$ set of feature vector and $-ve$ set of feature vector:

$$fp_i = \frac{fs_i}{\sum_{i=1}^{n_p} fs_i}, i = 1,2, \dots, n_p \quad \text{and} \quad fn_j = \frac{fs_j}{\sum_{j=1}^{n_n} fs_j}, j = 1,2, \dots, n_n \tag{6}$$

where, fp_i , fn_i are fuzzy membership values for $+ve$ and $-ve$ feature vectors respectively and n_p is number of $+ve$ feature vectors and n_n is that of $-ve$ feature vectors. Finally, we define the fuzzy membership based on nature of our problem is defined in eqn. (7) as:

$$\mu_i = \mu(x_i) = \begin{cases} P(+|x_i).\frac{\rho_i}{\rho}.fp_i\,, & y_i = +1 \ \ and \ \ , i = 1,2, \dots \dots, n_p \\ P(-|x_i).\frac{\rho_i}{\rho}.fn_i, & y_i = -1 \ \ and \ \ , i = 1,2, \dots \dots, n_n \end{cases} \tag{7}$$

The overall process of the fuzzy membership evaluation function is explained in Algorithm 1.

Algorithm 1. Fuzzy membership function evaluation

Input: We have the training set $= \{(x_1, y_1), (x_2, y_2), \dots \dots, (x_l, y_l)\}$, where $x_i \epsilon R^N$ belongs to one of the class for $y_i \epsilon(+1, -1)$ for $i = 1,2, \dots \dots, l$..

Output: The fuzzy membership μ_i , $i = 1,2, \dots \dots, l$.

Step 1 (a) Compute a 2-dimensional matrix R for distance of each vector to other vectors of set using (2).

(b) From the matrix R, find the maximum distance as:

$$d = \max_{i,j=1,2,\dots,j} d_{ij}$$

(c) Compute the average distance of all sample points using (3)

Note that the definition of sample density mainly uses the distance between the sample points of the original space. Because the specific form of $\sqrt{}(x)$ is unknown, the distance between the sample points in feature space can be obtained by kernel function $K(x, x')$

$$d\big(\phi(x), \phi(x')\big) = \|\phi(x) - \phi(x')\| = \sqrt{\big(\phi(x) - \phi(x')\big)^2}$$
$$= \sqrt{\phi(x)^2 - 2\big(\phi(x).\phi(x')\big) + \phi(x')^2}$$
$$= \sqrt{K(x,x) - 2.K(x,x') + K(x',x')}$$

The distance between any two samples points after mapping to the feature space can be found out through the above formulation.

Step 2 Compute the priori class probability $P\big(\omega_j\big)$, $j = 1,2$.

Step 3 Compute the class conditional probability $p\big(x_i|\omega_j\big)$.

Step 4 Compute the posterior probability $P\big(\omega_j|x_i\big)$.

Step 5 Compute the average density ρ and ρ_i for the sample set .

Step 6 Compute the fuzzy membership $\mu(x_i)$ using (4).

Step 7 Finally update the fuzzy membership μ_i using (5) to (7).

3 Results

For our analysis, we have used the Protein Data Bank (PDB) [10], and the Database of Interacting Proteins (DIP) [11], databases. The complete database, cross validation datasets and the source code for fuzzy SVM tool, developed under the current work is available freely to download for academic users from our website http://code.google.com/p/cmater-bioinfo/. The database involves 2008 positive interactions and 2408 negative interactions for *Homo sapiens* proteome [9]. It may be noted that the number of positive and negative interactions, considered in the experiment dataset for any proteome, are only a subset of all possible positive and negative interactions. This is done so, to limit the computational complexity of the training algorithm, during the multi-fold cross validation (CV) process. Each interacting or non-interacting residue fragments are represented using HQI8 amino acids indices [12] for both positive and negative data samples for the selected organism. Finally, we compare cross validation results with the fuzzy and classical SVM classifiers.

To analyze the performance of the developed technique, we have done a 10-fold cross validation experiment on the dataset, discussed above. In both the classical SVM and the fuzzy SVM classifier, we have used *polynomial* kernel function of degree 5 during experiments over the cross validation set. Ten cross validation experiment runs are marked as $run\#1, run\#2, \ldots, run\#10$. For each run of the experiment in both classical and fuzzy SVM training program, we vary three key kernel parameters $(c, \gamma \; and \; r)$ within a finite range. During any run of the cross validation experiment (run_i), the optimum set of kernel parameters are estimated as p_i [13, 14] and the best results in each run are reported in the results. The average cross validation performances using the classical SVM and fuzzy SVM classifiers on the three organisms are given in Table 1. In case of *Homo sapiens*, we see 1.55% of AUC improvement from classical SVM to F-SVM. As we know that MCC gives idea over the quality of binary classification and we have observed that there is significant improvement of MCC from classical SVM to F-SVM. The MCC gains of F-SVM over classical SVM classifier are 3.01% on *Homo sapiens* dataset (see Table 1). One of important classification parameter Sensitivity are also improved from classical SVM to F-SVM which are 2.14% on *Homo sapiens* dataset (see Table 1) respectively. However specificity shows lower rate of improvement whose respective gains are as 0.96% on *Homo sapiens* dataset (see Table 1). The test accuracy measure parameter, F-measure, is also improved by 1.55%, on *Homo sapiens* dataset (see Table 1).

Table 1. Comparison of average performances of 10 fold CV experiment over *Homo sapiens* data using classical SVM and fuzzy SVM classifiers

Classifier	Sensitivity	Specificity	MCC	F-measure	AUC
SVM	0.76495	0.80731	0.57285	0.76646	0.78613
F-SVM	0.78634	0.81687	0.60336	0.78392	0.80161

4 Conclusion

In the present work, we introduce the fuzzy SVM as a novel and accurate classifier for PPI site prediction with better performance than classical SVM classifier. We have

designed a new fuzzy membership function to give fuzzy value to each of the positive and negative fragment based on their interacting strength and with the help of the Bayesian formula also. We first transform the membership of F-SVM through posterior probability and weighted. It has been observed that through the simulation experiment in the same data set based on the posterior probability weighting membership of fuzzy support vector machine in this paper and it is better than classical SVM in classification AUC results. It shows a better performance over all the three organisms.

In this paper, we have worked with fuzzy classifier over Homo sapiens organism specific database. We would like to work it on other database by including more organisms in near future. Due to limitation of computing resources, all interactions could not be considered for CV experiment. Despite certain constraints, the current version of fuzzy SVM is observed to generate a steady and balanced prediction result over CV data set samples of the selected organisms. The fuzzy SVM classifier is also made available for free download in the public domain.

References

1. Korn, A., Burnett, R.: Distribution and complementarity of hydropathy in multi-subunit proteins. PROTEINS: Structure, Function, and Bioinformatics 9, 37–55 (1991)
2. Jones, S., Thornton, J.M.: Analysis of Protein-Protein Interaction Sites using Surface Patches. JMB 272, 121–132 (1997)
3. Lo Conte, L., Chothia, C., Janin, J.: The atomic structure of protein– protein recognition sites. J. Mol. Biol. 285, 2177–2198 (1999)
4. Jones, S., Thornton, J.M.: Principles of protein-protein interactions. Proceedings of the National Academy of Sciences 93(1), 13–20 (1996)
5. Lin, C.-F., Wang, S.-D.: Fuzzy Support Vector Machines. IEEE Transactions on Neural Networks 13(2) (2002)
6. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
7. Wei, Y., Wu, X.: A New Fuzzy SVM based on the Posterior Probability Weighting Membership. Journal of Computers 7(6), 1385–1392 (2012)
8. Sriwastava, B.K., Basu, S., Maulik, U., Plewczynski, D.: Prediction of E. coli Protein-Protein Interaction Sites Using Inter-Residue Distances and High-Quality-Index Features. In: Satapathy, S.C., Avadhani, P.S., Abraham, A. (eds.) Proceedings of the InConINDIA 2012. AISC, vol. 132, pp. 837–844. Springer, Heidelberg (2012)
9. Sriwastava, B.K., Basu, S., Maulik, U., et al.: PPIcons: identification of protein-protein interaction sites in selected organisms. Journal of Molecular Modeling, 1–12 (2013)
10. Berman, H., Westbrook, J., Feng, Z., et al.: The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000)
11. Salwinski, L., Miller, C.S., Smith, A.J., et al.: The Database of Interacting Proteins: 2004 update. Nucleic Acids Research 32, D449–D451 (2004)
12. Saha, I., Maulik, U., Bandyopadhyay, S., et al.: Fuzzy Clustering of Physicochemical and Biochemical Properties of Amino Acids. Amino Acids (2011)
13. Basu, S., Plewczynski, D.: AMS3.0: prediction of post-translational modifications. BMC Bioinformatics 11, 210 (2010)
14. Plewczynski, D., Basu, S., Saha, I.: AMS 4.0: consensus prediction of post-translational modifications in protein sequences. Amino Acids 43(2), 573–582 (2012)