hany.alashwal@gmail.com

One-Class Classifier to Predict Protein-Protein Interactions based on Hydrophobibity Properties

Hany Alashwal, Safaai Deris, Razib M. Othman, and Mohd S. Mohamad

Abstract— **Protein-protein interactions are important in a wide range of biological processes. The development of drugs that target such interactions is a very active research field. Hence predicting protein-protein interactions represent an important challenge in bioinformatics research. Machine learning techniques have been applied to predict protein-protein interactions. Most of these techniques address this problem as a binary classification problem. While it is easy to get a dataset of interacting protein as positive example, there are no experimentally confirmed noninteracting proteins to be considered as a negative set.**

Therefore, in this paper we solve this problem as a one-class classification problem using One-Class SVM (OCSVM). The hydrophobicity properties have been used in this research as the protein sequence feature.

Using only positive examples (interacting protein pairs) for training, the OCSVM achieves accuracy of 72% using RBF kernel. These results imply that protein-protein interaction can be predicted using oneclass classifier with reliable accuracy.

I. Introduction

THE recent studies of molecular biology led the researchers to recognize that protein-protein interactions affect almost all processes in a cell [1], [2]. In the last few years, the problem of computationally predicting protein-protein interactions has gain a lot of attention. It has been shown that proteins with similar functions are more likely to interact [2]. If the function of one protein is known then the function of its binding partners is likely to be related. This helps to understand the functional roles of unannotated protein by knowing its interaction partners. Drug discovery is another area where protein–protein interaction prediction plays an important role.

For that reasons, identifying protein-protein interactions represents a crucial step toward understanding proteins functions. In the last few years, the problem of computationally predicting proteinprotein interactions has gain a lot of attention. Methods based on the machine learning theory have been proposed [3]-[5]. Most of these methods address this problem as a binary classification problem. Although, constructing a positive dataset (i.e. pairs of interacting proteins) is relatively an easy task by using one of the available databases of interacting proteins, there is no data on experimentally confirmed non-interacting protein pairs have been made available. To cope with this problem, some researchers created an artificial negative protein interaction dataset for S. cerevisiae by randomly generating 100,000 protein pairs from this organism that are not described as interacting in the Database of Interacting Proteins (DIP) [6] without putting any further restrictions on such pairs, as in [5].

However, since only data of interacting proteins pairs (positive data) are available and sampled well, the problem of predicting protein-protein interactions is fundamentally a one class classification problem. In this respect, we propose a recent method, one-class support vector machines (OCSVMs) for proteinprotein interactions predictions.

II. Data Set and Features Representation

The protein interaction data was obtained from the Database of Interacting Proteins (DIP) [6]. The DIP database was developed to store and organize information on binary protein–protein interactions that was retrieved from individual research articles. The DIP database provides sets of manually compiled protein-protein interactions in Saccharomyces cerevisiae. The majority of DIP entries are obtained from combined, non-overlapping data mostly obtained by systematic two-hybrid analyses. The current version contains 4749 proteins involved in 15675 interactions for which there is domain information. DIP also provides a high quality core set of 2609 yeast proteins that are involved in 6355 interactions which have been determined by at least one small-scale experiment or at least two independent experiments and predicted as positive by a scoring system [6].

The proteins sequences files were obtained for the Saccharomyces Genome Database (SGD) [7]. The SGD project collects information and maintains a database of the molecular biology of the yeast Saccharomyces cerevisiae. This database includes a variety of genomic and biological information and is maintained and updated by SGD curators. The proteins sequence information is needed in this research in order to elucidate the domain structure of the proteins involved in the interaction and to represent the amino acid hydrophobicity in the feature vectors. The construction of an appropriate feature space that describes the training data is essential for any supervised machine learning system. The amino acid hydrophobicity properties can be used to construct the feature vectors for SVM. The amino acids hydrophobicity

Authors are with Artificial Intelligence and Bioinformatics Laboratory, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia.

properties are obtained from [8]. The hydrophobicity features can be represented in feature vector as:

$$\mathbf{x} = [h_1, h_2, \ldots, h_i, \ldots, h_n] \qquad (1)$$

where k is the number of amino acid in the protein x, h_i= 1 when the amino acid is hydrophobic and h_i = 0 when the amino acid is hydrophilic.

III. One-Class Support Vector Machines

One-class classification problem is a special case from the binary classification problem where only data from one class are available and sampled well. This class is called the target class. The other class which is called the outlier class, can be sampled very sparsely, or can be totally absent. It might be that the outlier class is very hard to measure, or it might be very expensive to do the measurements on these types of objects. For example, in a machine monitoring system where the current condition of a machine is examined, an alarm is raised when the machine shows a problem. Measurements on the normal working conditions of a machine are very cheap and easy to obtain. On the other hand, measurements of outliers would require the destruction of the machine in all possible ways. It is very expensive, if not impossible, to generate all faulty situations. Only a method trained on just the target data can solve the monitoring problem.

Basically, one-class SVM treats the origin as the only member of the second class (see Fig. 1). Then using relaxation parameters, it separates the members of the one class from the origin. Then the standard binary SVM techniques are employed.

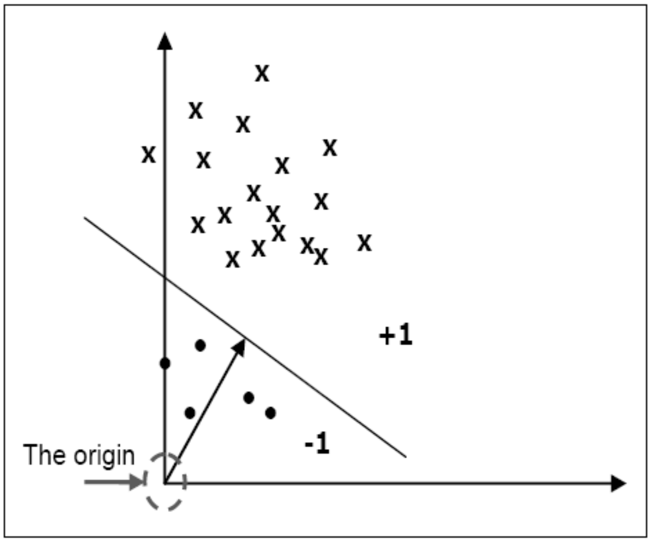


Fig. 1. Classification in one-class SVM

The OCSVM algorithm maps input data into a high dimensional feature space (via a kernel) and iteratively finds the maximal margin hyperplane which best separates the training data from the origin. The OCSVM may be viewed as a regular two-class SVM where all the training data lies in the first class, and the origin is taken as the only member of the second class. Thus, the hyperplane (or linear decision boundary) corresponds to the classification function:

$$f(x) = \langle w, x \rangle + b \qquad (2)$$

where w is the normal vector and b is a bias term. The OCSVM solves an optimization problem to find the function f with maximal geometric margin. We can use this classification function to assign a label to a test example x. If f (x) < 0 we label x as an anomaly, otherwise it is labeled normal.

Using kernels, solving the OCSVM optimization problem is equivalent to solving the following dual quadratic programming problem:

$$\min_{\alpha} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \qquad (3)$$

Subject to $0 \le \alpha_i \le \dfrac{1}{vl},\ and\ \sum_i \alpha_i = 1$ (4)

where α_i is a Lagrange multiplier (or "weight" on example i such that vectors associated with non-zero weights are called "support vectors" and solely determine the optimal hyperplane), ν (*nu*), is a parameter that controls the trade-off between maximizing the distance of the hyperplane from the origin and the number of data points contained by the hyperplane, l is the number of points in the training dataset, and K (x_i ,x_j) is the kernel function. By using the kernel function to project input vectors into a feature space, we allow for nonlinear decision boundaries. Given a feature map:

$$\phi : X \rightarrow \Re^N \qquad (5)$$

where φ maps training vectors from input space X to a high-dimensional feature space, we can define the kernel function as:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \qquad (6)$$

Feature vectors need not to be computed explicitly, and in fact it greatly improves computational efficiency to directly compute kernel values $K(x_i, x_j)$.

IV. Results and Discussion

We developed programs using Perl for parsing the DIP databases, sampling of records and sequences, and replacing amino acid sequences of interacting proteins with its corresponding feature. To make a positive interaction set, we represent an interaction pair by concatenating feature vectors of each proteins pair that are listed in the DIP-CORE as interacting proteins. Since we use domain feature we include only the proteins that have structure domains. The resulting positive set for domain feature contains 1879 protein pairs.

In our computational experiment, we employed the LIBSVM (version 2.5) software and modified it to train and test the one-class SVM proposed in this paper. This is an integrated software tool for support vector classification, regression, and distribution estimation, which can handle one-class SVM. In order to train our one-class SVMs, we examine out the following four kernels find appropriate parameter values:

- Linear: $K(x_i,x_j)=x_i^T x_j$.
- Polynomial: $K(x_i,x_j) = (\gamma x_i^T x_j + r)^d , \gamma > 0$.
- Radial basis Function (RBF): $K(x_i, x_j) = \exp(\gamma \| x_i - x_j \|^2), \gamma > 0$.
- Sigmoid: $K(x_i,x_j) = tahn(\gamma x_i^T x_j + r)$.

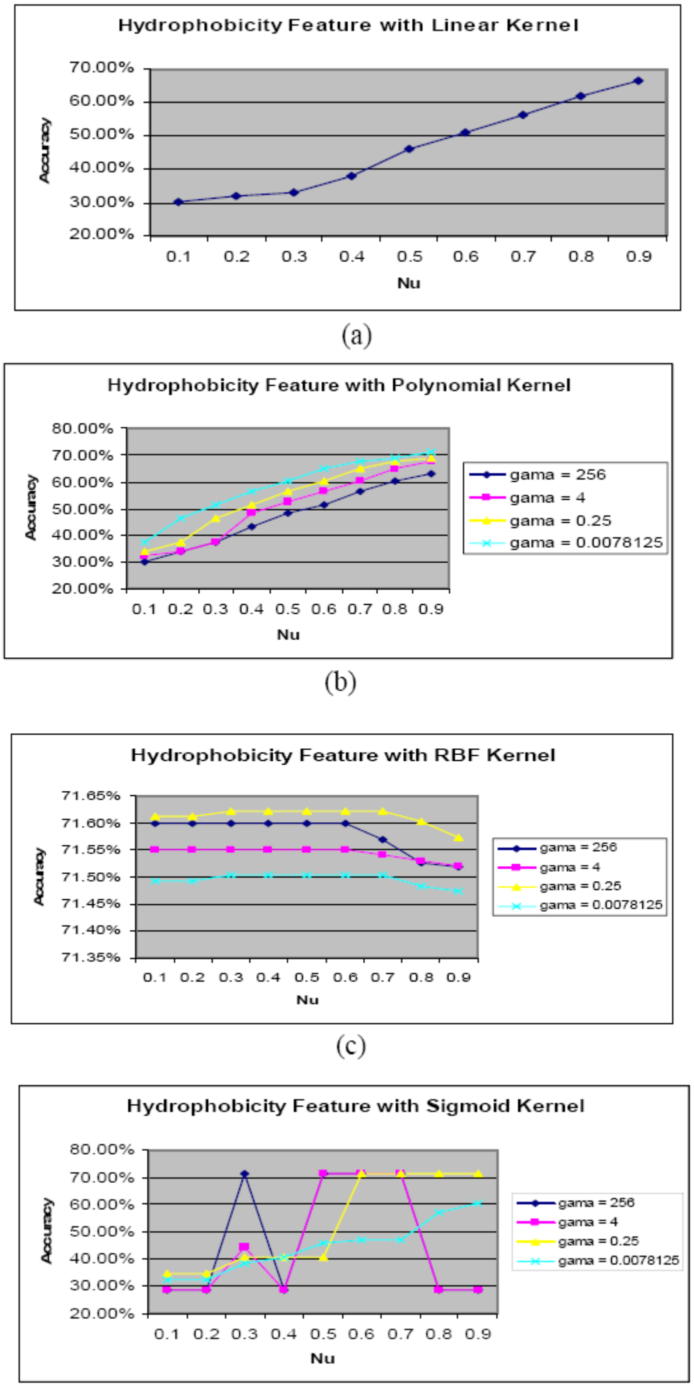
where γ (*gama*), *r*, and *d* are kernel parameters to be set for a specific problem. We carried out our experiments using the above mentioned kernels.

The results of our experiments are summarized in Fig. 3. These results indicate that it is informative enough to consider the hydrophobicity properties of the amino acids in the protein pairs to facilitate the prediction of protein-protein interactions. These results also indicate that the difference between interacting and non-interacting protein pairs can be learned from the available data using one-class classifier. It is also important to note that the choice of the parameters has a clear impact on the classifier performance.

Appropriate parameters for one-class SVMs with different kernels are set by the cross-validation process. We can see from this validation process that it is important to choose the appropriate parameters. As shown in Figure 2, OCSVM is very sensitive to the choice of parameters. However, since one-class SVMs with linear kernel does not have the parameter *gama*, we executed the cross-validation process only for parameter *nu*. Then the cross-validation accuracy is calculated in each run as the number of corrected prediction divided by the total number of data ((TP+TN)/(TP+FP+TN+FP)). Then the average is calculated for the 10 folds.

The best results were found by the Sigmoid kernel (Fig. 2 (d)). Even though, the Sigmoid kernel could give as low accuracy as 29% with unsuitable choice of parameters, it achieves around 72% with proper choice of parameters. However, the RBF kernel is the most stable kernel that also gives a comparable accuracy.

In addition, the one-class SVM approach is better reflecting the reality of the problem than the conventional binary classifiers. This is due to the fact that all the binary classifiers need to be trained using two classes. In contrast, one-class SVM is able to work, solely on the basis of target examples. Therefore one-class SVM has the advantage of using only real data in the training phase.



(a)

(b)

(c)

(d)

Fig. 2. One-class SVM performance for proteins interactions using different kernels

V. Conclusion

The problem of predicting protein-protein interactions possesses the features of one-class classification problem where only data from target class (i.e. interacting proteins) are available and sampled well. Therefore, the objective of this paper was to show that the one-class SVM method can be applied successfully to the problem of predicting protein-protein interactions. Experiments performed on real dataset show that the performance of this method is comparable to that of normal binary SVM using artificially generated negative set. Of course, the absence of negative information entails a price, and one should not expect as good results as when they are available. In conclusion the result of this study suggests that protein-protein interactions can be predicted from domain structure with reliable accuracy. Consequently, these results show the possibility of proceeding directly from the automated identification of a cell's gene products to inference of the protein interaction pairs, facilitating protein function and cellular signaling pathway identification.

References

[1] H. Lodish, A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular cell biology* (4th edition). W.H. Freeman, New York, 2000.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K.Roberts, and P. Walter, *Molecular Biology of the Cell* (4th edition). Garland Science, 2002.

[3] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17(5), pp: 455-460, 2001.

[4] Y. Chung, G. Kim, Y. Hwang, and H. Park, "Predicting Protein-Protein Interactions from One Feature Using SVM," *In proceedings of IEA/AIE*'04, pp:50-55, 2004.

[5] S. Dohkan, A. Koike and T. Takagi, "Prediction of protein-protein interactions using Support Vector Machines," *In Proceedings of the Fourth IEEE Symposium on BioInformatics and BioEngineering*

[6] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30(1), pp: 303- 305, 2002.

[7] E. L. Hong, R. Balakrishnan, K.R. Christie, M.C. Costanzo, S.S. Dwight, S.R. Engel, D.G. Fisk, et al., "Saccharomyces Genome Database" http://www.yeastgenome.org/, (25th Dec 2005).

[8] T. P. Hopp and K. R. Woods, "Predicting of protein antigenic determinants from amino acid sequences," Proc. Natl Acad. Sci. USA, 78, 3824-3828, 1981.MillerE H 2003 *IEEE Trans. Antennas Propagat.*., to bepublished.