Prediction of Protein Subcellular Multi-locations with a Min-Max Modular Support Vector Machine

Yang Yang and Bao-Liang Lu*

Department of Computer Science and Engineering,
Shanghai Jiao Tong University,
800 Dong Chuan Rd., Shanghai 200240, China
{alayman, bllu}@sjtu.edu.cn

Abstract. How to predict subcellular multi-locations of proteins with machine learning techniques is a challenging problem in computational biology community. Regarding the protein multi-location problem as a multi-label pattern classification problem, we propose a new predicting method for dealing with the protein subcellular localization problem in this paper. Two key points of the proposed method are to divide a seriously unbalanced multi-location problem into a number of more balanced two-class subproblems by using the part-versus-part task decomposition approach, and learn all of the subproblems by using the min-max modular support vector machine (M^3-SVM). To evaluate the effectiveness of the proposed method, we perform experiments on yeast protein data set by using two kinds of task decomposition strategies and three kinds of feature extraction methods. The experimental results demonstrate that our method achieves the highest prediction accuracy, which is much better than that obtained by the existing approach based on the traditional support vector machine.

1 Introduction

The localization of a protein in a cell is very important for understanding its function. Due to the difficulties of conducting biological experiments to determine the subcellular locations, a lot of efforts have been made to develop automatic tools for localization. As the numbers of new genome and protein sequences in the public databases have increased dramatically in recent years, methods based on analyzing protein sequences have been largely developed. In 1994, Nakashima and Nishikawa discriminated intracellular and extracellular proteins successfully by amino acid composition and residue-pair frequencies [1]. Till now, many locations have been successfully discriminated and various pattern classification and machine learning methods have been used, such as Mahalanobis distance

* To whome correspondence should be addressed. This work was supported by the National Natural Science Foundation of China under the grants NSFC 60375022 and NSFC 60473040.

J. Wang et al. (Eds.): ISNN 2006, LNCS 3973, pp. 667–673, 2006.

[2], neural network [3], hidden Markov model (HMM) [4] and support vector machine [5].

Most of these researches focus on mono-locational proteins, i.e., proteins existing in only one location. However, a lot of proteins bear multi-locational characteristics. According to our statistics of Swiss-Prot database [8], there are more than five thousands proteins locating in more than one location. Recently, Cai and Chou first tackled the classification of multi-locational proteins in yeast [10]. They used GO-FunD-PseAA method, which hybridizes gene ontology, functional domain composition and pseudo-amino acid composition approach. Although this method improves the prediction accuracy a lot, it fails to give a general classification method for this multi-location problem. In addition, there are a large portion of proteins lack the information like GO and FunD.

In this paper, we apply M^3-SVM to solve the problem. Several feature extraction methods are also discussed, including amino acid composition, amino acid pair composition and segmentation method. A series of standard measures are used to evaluate the classification performance. The experimental results show that using M^3-SVM and the part-versus-part strategy can get a much higher prediction accuracy than traditional SVM and other classification methods.

2 Our Method

2.1 Min-Max Modular Support Vector Machine

The min-max modular network has been shown to be an efficient classifier, especially in solving large-scale and complex multi-class pattern classification problems [6]. It divides a complex classification problem into many small independent two-class classification problems, which can be learned parallelly without communication with each other. And then it integrates these modules to get a final solution to the original problem according to two module combination rules, namely minimization and maximization principles. The min-max modular support vector machine [7], which use SVM as base classifier and M^3 network structure, has been successfully used in many pattern classification problems, such as text categorization, human face recognition and industrial fault image detection.

2.2 Part-Versus-Part Strategy

As for multi-class problems, one-versus-rest decomposition is usually used [9]. Given a K-class multi-label problem, its training set is as follows:

$$\mathcal{X} = \{(x_m, t_m)\}_{m=1}^{l}, t_m = \{t_m^k\}, k = 1, ..., \tau_m \tag{1}$$

where $x_m \in \mathrm{R}^n$ is the mth sample in the data set, t_m is the label set of x_m, t_m^k is the kth label of x_m, and τ_m denotes the number of labels of x_m.

Decompose the K-class multi-label problem T to K two-class problems $T_i, i = 1, ..., K$. The training set of T_i is defined as

$$\mathcal{X}_i = \{(x_m^{i+}, +1)\}_{m=1}^{l_i^+} \cup \{(x_m^{i-}, -1)\}_{m=1}^{l_i^-} \tag{2}$$

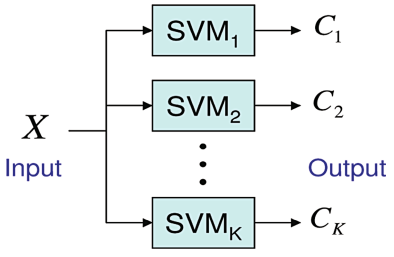

Fig. 1. A multi-label problem divided into several two-class subproblems with the one-versus-rest strategy

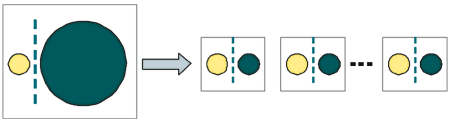

Fig. 2. Decomposition of a seriously unbalanced multi-lable problem into a number of balanced two-class subproblems with the part-versus-part strategy for M^3-SVM

where l_i^+ is the number of positive samples of the two-class problem T_i, and l_i^- is the number of the negative samples. For T_i, positive samples are those whose label sets contain the label C_i and negative samples are the remaining ones. Figure 1 depicts a multi-label problem divided into several two-class subproblems with the one-versus-rest strategy and SVM as the two-class classifier.

Considering that many biological problems have unbalanced data distribution for the classes, such as proteins occurring in Cytoplasmic, Nuclear and Plasma membrane being much more than those in other locations, we adopt part-versus-part strategy here [7]. An important advantage of the part-versus-part method over existing popular pairwise-classification approach is that a large-scale two-class subproblem can be further divided into a number of relatively smaller and balanced two-class subproblems, and fast training of SVMs on massive multi-class classification problems can be easily implemented in a massively parallel way.

The part-versus-part decomposition is straightforward which further decomposes the two-class problems to smaller ones as shown in Figure 2. For a two-class problem T_i, its positive and negative training set $\mathcal{X}_i^+$ and $\mathcal{X}_i^-$ can be further decomposed into N_i^+ and N_i^- subsets, where $1 \le N_i^+ \le l_i^+$, $1 \le N_i^- \le l_i^-$.

$$\mathcal{X}_{ij}^+ = \{(x_m^+, +1)\}_{m=1}^{l_{ij}^+}, j = 1, ..., N_i^+ \tag{3}$$

$$\mathcal{X}_{ij}^- = \{(x_m^-, -1)\}_{m=1}^{l_{ij}^-}, j = 1, ..., N_i^- \tag{4}$$

The l_{ij}^+ and l_{ij}^- are numbers of samples in $\mathcal{X}_{ij}^+$ and $\mathcal{X}_{ij}^-$, respectively.

After the original problem is divided into related balanced subproblems, each of which can be handled by a SVM. And We use min-max modular network to organize all the subproblem together.

2.3 Task Decomposition

Task decomposition is a key problem for M^3-SVM. A good decomposition method can maintain or even improve the generalization performance. In this paper, we use two kinds of methods, namely random decomposition and hyperplane decomposition [13]. The former is simple and straightforward. Given a

specific module size, it chooses samples randomly from the training set to form a new smaller training set. This method can not obtain a stable performance and may hurt the generalization ability sometimes.

As for the hyperplane decomposition method, a series of specific hyperplanes are introduced and the training data are sorted according to their distances to the hyperplanes. Then the ordered sequence of training data will be divided into relatively balanced subsets.

3 Results and Discussion

3.1 Data Set

We conducted experiments on a data set collected from Swiss-Prot according to the list of codes of the 4,709 budding yeast proteins given in [10]. None of the proteins included here has 40% sequence identity with any other. Because some sequences are absent in the database, the data set we used is 19 ones less than theirs. But it would not has much impact on the overall accuracy. The distribution of the data set is listed in Tables 1 and 2. We adopted 10-fold cross-validation test. All experiments were performed on a 3GHz Pentium 4 PC with 2GB RAM.

Table 1. Numbers of proteins for every class

Location	Sequence No.	Location	Sequence No.
Actin	29	Lipid particle	19
Bud	23	Microtubule	20
Bud neck	59	Mitochondrion	491
Cell periphery	104	Nuclear periphery	59
Cytoplasm	1565	Nucleolus	156
Early Golgi	51	Nucleus	1323
Endosome	43	Peroxisome	20
ER	271	Punctuate composite	123
ER to Golgi	6	Spindle pole	58
Golgi	40	Vacuolar membrane	54
Late Golgi	36	Vacuole	129
Summation of all classes		4679	
Number of different proteins		3536	

Table 2. Distribution of multi-locational proteins

Number of Locations	1	2	3	4	5
Number of Sequences	2465	1007	57	6	1

3.2 Experimental Results

A proper representation for protein sequences is very important to the classification of proteins. Researchers have developed a lot of features extraction methods

for protein sequences. Here we experimented three approaches: amino acid composition (AAC), amino acid pair composition (AAP) and segmentation method (SEG)[12]. Each protein in the data set of l proteins will be characterized by a vector $v_i (i = 1, \cdots, l)$, which represents sequence features.

AAC is a conventional method which converts a protein sequence $\mathcal{S}$ to a vector $v = \{a_1, a_2, \ldots, a_{20}\}$, where $a_i (1 \leq i \leq 20)$ reflects the occurrence frequency of one of the 20 amino acids ({A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}) in a protein. AAP contains 400 components, each representing an amino acid pair composition by counting two adjacent amino acids overlappingly. The SEG method regards protein sequences as text and segment them nonoverlappingly by match words in an established dictionary. The length of words used for segmentation is not limited to two but determined according to need. Moreover, it does not use all the k-mers but select informative ones by some criteria. Here we establish a dictionary of 30 words, including 20 amino acid, 5 most frequent amino acid pair and 5 3-kmers. The SEG method performs the best with traditional SVM. All of the three methods can obtain better prediction accuracy using M^3-SVM.

To evaluate the effectiveness of the multi-label classification comprehensively, we use recall, precision and F_1 measure for each class. We trained the classifier with a RBF kernel and set the module size of M^3-SVM to 100. Since the task

Table 3. Results by using SVM and M^3-SVM

Location	M^3-SVM(R)			M^3-SVM(H)			SVM		
	R	P	F_1	R	P	F_1	R	P	F_1
Actin	13.8	4.1	6.3	17.2	13.9	**15.4**	0.0	0.0	0.0
Bud	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bud neck	76.3	3.5	6.7	54.2	5.3	**9.7**	1.7	33.3	3.2
Cell periphery	15.4	6.6	9.2	26.9	7.5	**11.7**	1.0	9.1	1.7
Cytoplasm	96.3	46.4	62.4	87.5	46.8	61.0	80.3	56.2	**66.3**
Early Golgi	33.3	7.8	12.6	35.3	7.7	**12.6**	0.0	0.0	0.0
Endosome	9.3	2.5	3.9	20.9	2.8	**5.0**	0.0	0.0	0.0
Endoplasmic reticulum	52.8	23.4	32.4	64.9	20.8	31.5	33.9	45.1	**38.7**
ER to Golgi	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Golgi	10.0	2.9	4.5	5.0	3.9	4.4	2.5	50.0	**4.8**
Late Golgi	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lipid particle	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Microtubule	5.0	2.0	**2.8**	0.0	0.0	0.0	0.0	0.0	0.0
Mitochondrion	75.4	23.9	36.3	69.5	31.8	43.7	40.5	55.3	**46.8**
Nuclear periphery	6.8	2.7	3.8	27.1	5.3	8.8	8.5	38.5	**13.9**
Nucleolus	61.5	8.7	15.2	58.3	9.8	**16.8**	1.3	28.6	2.5
Nucleus	46.3	44.5	45.4	86.3	40.9	**55.5**	41.2	57.1	47.8
Peroxisome	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Punctate composite	12.2	3.0	4.8	54.5	4.7	**8.7**	0.0	0.0	0.0
Spindle pole	31.0	10.3	15.5	51.7	10.0	**16.7**	1.7	100.0	3.4
Vacuolar membrane	11.1	4.0	5.9	14.8	5.7	**8.2**	0.0	0.0	0.0
Vacuole	41.1	11.9	**18.4**	63.6	10.7	18.3	3.9	17.9	6.4

decomposition is one of the two key problems of M^3-SVM, two kinds of task decomposition methods were experimented. One is the random task decomposition strategy, the other is hyperplane task decomposition strategy. The detailed values of recall, precision, F_1 of 22 classes are given in Tables 3. Here amino acid composition method is adopted. Let M^3-SVM(R) stand for M^3-SVM with the random strategy, and M^3-SVM(H) the hyperplane strategy.

From the experimental results, we can observe that M^3-SVM(H) performs the best among the three methods. And many small classes were successfully discriminated by using M^3-SVM with part-versus-part decomposition, while SVM classified all the proteins to several big classes.

3.3 Comparison with Other Methods

Chou and Cai has reported that the likelihood of hitting the localization of a protein in budding yeast could be as high as 90% [14] using GO-FunD-PseAA method. In their method, gene ontology and functional domain knowledge are used for prediction. Since we aim to propose a general classification method, we make comparisons with other methods based on the same feature vectors, i.e., the amino acid composition. The Least Euclidean Distance algorithm, Least Hamming Distance algorithm and ProtLoc predictor obtained success rates of 13.89, 14.03 and 13.95%, respectively [10]. According to our experimental results, traditional SVM obtained overall success rate of 46%. The M^3-SVM(H) and M^3-SVM(R) obtained accuracies of 73% and 64%, respectively, which are much higher than other classification methods.

4 Conclusions and Future Work

This study focuses on seeking efficient classification method to predict subcellular locations for proteins existing in one or more locations. We apply M^3-SVM and part-versus-part strategy to solve this multi-label problem. And several feature extraction methods for protein sequences are compared. The experiments were conducted on a data set of yeast proteins. Results show that the classification method we proposed is superior to other methods on a series of performance measures and improves the accuracy significantly.

As a future work, we will consider referring other available field knowledge to get more precise prediction results. And now we are constructing large-scale data sets covering various species from Swiss-Prot. We believe that our methods will be competent in solving new complex classification tasks.

Acknowledgements

The authors thank Mr. Ken Chen for his helpful work on data preparation.

References

1. Nakashima, H., Nishikawa, K.: Discrimination of Intracellular and Extracellular Proteins Using Amino Acid Composition and Residue-pair Frequencies. J. Mol. Biol. 238 (1994) 54-61
2. Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E.: Relation Between Amino Acid Composition and Cellular Location of Proteins. J. Mol. Biol. 266 (1997) 594-600
3. Reinhardt, A., Hubbard, T.: Using Neural Networks for Prediction of the Subcellular Location of Proteins. Nucleic Acids Research 26 (1998) 2230-2236
4. Fujiwara, Y., Asogawa, M., Nakai, K.: Prediction of Mitochondrial Targeting Signals Using Hidden Markov Models. Genome Informatics (1997) 53-60
5. Hua, S., Sun, Z.: Support Vector Machine Approach for Protein Subcellular Localization Prediction. Bioinformatics 17 (2001) 721-728
6. Lu, B.L., Ito, M.: Task Decomposition and Module Combination Based on Class Relations: a Modular Neural Network for Pattern Classification. IEEE Transactions on Neural Networks 10 (1999) 1244-1256
7. Lu, B.L., Wang, K.A., Utiyama, M., Isahara, H.: A Part-Versus-Part Method for Massively Parallel Training of Support Vector Machines. Proceedings of International Joint Conference on Neural Networks (2005) 735-740
8. Chen, K., Liang, W.M., Lu, B.L.: Data Analysis of Swiss-Prot Database. BCMI Technical Report BCMI-TR-0501, Shanghai Jiao Tong University (2005)
9. Joachims, T.: Learning to Classify Text Using Support Vector Machine: Method, Theory, and Algorithms. Kluwer Academic Publishers (2002)
10. Chou, K.C., Cai, Y.D.: Prediction of Protein Subcellular Locations by GO-FunD-PseAA Predictor. Biochemical and Biophysical Research Communications 320 (2004) 1236-1239
11. Apweiler, R.: The InterPro Database, an Integrated Documentation Resource for Protein Families, Domains and Functional Sites. Nucleic Acids Research 29 (2001) 37-40
12. Yang, Y., Lu, B.L.: Extracting Features from Protein Sequences Using Chinese Segmentation Techniques for Subcellular Localization. Proceedings of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (2005) 288-295
13. Liu, F.Y., Wu, K., Zhao, H., Lu, B.L.: Fast Text Categorization with Min-Max Modular Support Vector Machines. Proceedings of International Joint Conference on Neural Networks (2005) 570-575
14. Chou, K.C., Cai, Y.D.: Predicting Protein Localization in Budding Yeast. Bioinformatics 21(7) (2005) 944-950