Clustering Support Vector Machines and Its Application to Local Protein Tertiary Structure Prediction

Jieyue He[1], Wei Zhong[2], Robert Harrison[2,3,4] Phang C. Tai[3,] and Yi Pan[2,]*

[1] Department of Computer Science
Southeast University, Nanjing 210096, China
jieyuehe@seu.edu.cn
[2] Department of Computer Science,
[3] Department of Biology
Georgia State University, Atlanta, GA 30303-4110, USA
[4] GCC Distinguished Cancer Scholar
pan@cs.gsu.edu

Abstract. Support Vector Machines (SVMs) are new generation of machine learning techniques and have shown strong generalization capability for many data mining tasks. SVMs can handle nonlinear classification by implicitly mapping input samples from the input feature space into another high dimensional feature space with a nonlinear kernel function. However, SVMs are not favorable for huge datasets with over millions of samples. Granular computing decomposes information in the form of some aggregates and solves the targeted problems in each granule. Therefore, we propose a novel computational model called Clustering Support Vector Machines (CSVMs) to deal with the complex classification problems for huge datasets. Taking advantage of both theory of granular computing and advanced statistical learning methodology, CSVMs are built specifically for each information granule partitioned intelligently by the clustering algorithm. This feature makes learning tasks for each CSVMs more specific and simpler. Moreover, CSVMs built particularly for each granule can be easily parallelized so that CSVMs can be used to handle huge datasets efficiently. The CSVMs model is used for predicting local protein tertiary structure. Compared with the conventional clustering method, the prediction accuracy for local protein tertiary structure has been improved noticeably when the new CSVM model is used. The encouraging experimental results indicate that our new computational model opens a new way to solve the complex classification for huge datasets.

1 Introduction

Support Vector Machines (SVMs) [18] are new generation of machine learning techniques and have been successfully applied to a wide variety of application domains [8] including bioinformatics [15]. SVMs are searching the optimal separating hyperplane by solving a convex quadratic programming (QP). The typical running time for

* Corresponding author.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, pp. 710 – 717, 2006.

the convex quadratic programming is $\Omega(m^2)$ for the training set with m samples. The convex quadratic programming is NP-complete in the worst case [19]. Therefore, SVMs are not favorable for a large dataset [7]. In the local protein tertiary structure prediction, our dataset contains half millions samples. According to [23], it would take years to train SVMs on a dataset containing one million records.

Many algorithms and implementation techniques have been developed to enhance SVMs in order to increase their training performance with large data sets. The most well-known techniques include chunking [16], Osuna's decomposition method [13], and Sequential Minimal Optimization (SMO)[14]. The success of these methods depends on dividing the original Quadratic Programming (QP) problem into a series of smaller computational problems in order to reduce the size of each QP problem. Although these algorithms accelerate the training process, these algorithms do not scale well with the size of the training data.

The second class of algorithms tries to speed up the training process by reducing the number of training data. Since some data points such as the support vectors are more important to determine the optimal solution, these algorithms provide SVMs with high quality data points during the training process. Random Selection [1,3], Bagging [17], and clustering analysis [2, 9, 23] are representatives of these algorithms. These algorithms are highly scalable for the large dataset while the performance of training depends greatly on the selection of training samples.

Unlike those two classes of methods to deal with huge dataset training, we propose a new computational model called Clustering Support Vector Machines Model (CSVMs Model) by combining the theory of granular computing and principles of the statistical learning algorithms. The difference between the CSVMs model and clustering analysis [2, 9, 23] is that all the training dataset is kept during the training process. Using the principle of granular computing, the CSVMs model is able to divide a complex data-mining problem into a series of smaller and computationally simpler problems [22].

Han and Baker have used the conventional clustering algorithm to predict local protein structure [5, 6]. In their work, the K-means clustering algorithm is essential to understand how protein sequences correspond to local protein tertiary structures. Since K-means clustering algorithm may not reveal the nonlinear sequence-to-structure relationship effectively, many sequence segments in one cluster are weakly mapped to the representative structure of their assigned clusters. Consequently, the structure prediction quality for these clusters is not satisfactory.

In order to overcome the problems of the conventional clustering algorithm, the CSVMs model is used to predict local protein structure in this study. In this new computational model, one SVM is built particularly for each information granule defined by sequence clusters created by the clustering algorithm. CSVMs are modeled to learn the nonlinear relationship between protein sequences and their structures in each cluster. SVM is not favorable for large amount of training data samples. However, CSVMs can be easily parallelized to speed up the modeling process. After gaining the knowledge about the sequence to structure relationship, CSVMs are used to predict the local tertiary structure for protein sequence segments. Local tertiary structure is represented by distance matrices, torsion angles and secondary structures for backbone α-carbon atoms of protein sequence segments. In this study, the performance of the CSVM

model and the conventional clustering algorithm are compared based on accuracy for local protein structure prediction.

The paper is organized as follows. Section 2 describes a new computational model called Clustering Support Vector Machines Model. Section 3 explains the experimental setup and result analysis. Finally, the conclusion and the future work are presented in section 4.

2 Clustering Support Vector Machines Model (CSVM Model)

Model construction and performance evaluation are two major tasks for Clustering Support Vector Machines Model (CSVMs Model). In the model construction phase, the whole sequence space is granulated into a series of information granule and each CSVM is built specifically for each information granule. In the performance evaluation phase, the established CSVM model is used to predict local tertiary structures for sequence segments.

2.1 Granulating the Whole Sequence Space into Clusters in the Model Construction Phase

Granular computing decomposes information in the form of some aggregates and solves the targeted problems in each granule [21]. Fuzzy sets, probabilistic sets, decision trees, clusters and association rules are some of granulation methods under the framework of granular computing [22]. Since K-means clustering is computationally efficient for large datasets [10], K-means clustering is chosen as the granulation method in our study. With the K-means clustering algorithm, data samples with similar characteristics can be grouped together. As a result, the whole sample space is partitioned into subspaces intelligently and the complex data mining work is mapped into a series of computationally tractable tasks. Different number of initial clusters were tried and based on these results, 800 clusters were chosen. 800 clusters are relatively suitable for the K-means clustering algorithm in our application.

2.2 Building Clustering Support Vector Machines (CSVMs) for Each Cluster in the Model Construction Phase

Because the distribution patterns for frequency profiles in clusters are diverse, the functionality of CSVMs is customized for each cluster belonging to different cluster groups. The definition of different cluster groups is introduced in the section explaining the experimental setup. The CSVMs for clusters belonging to the bad cluster group are designed to select potentially reliable prediction since a large percentage of prediction by the bad cluster group is not accurate. The CSVMs for clusters belonging to the good cluster group are designed to filter out potentially unreliable prediction since the large percentage of prediction by the good cluster group is highly reliable.

Since CSVM solves the binary classification problem, samples in a cluster must be labeled as positive or negative before the training process. In each cluster, positive samples are defined as those samples whose structure deviation from the representative structure of this cluster is within a given threshold. Negative samples are similarly defined. Positive samples have the potential to be closely mapped to the

representative structure of the specified cluster and negative samples may not correspond to the representative structure of the specified cluster closely. Each CSVM is trained specifically for one cluster.

2.3 Local Protein Tertiary Structure Prediction by CSVMs Model

Local protein tertiary structure prediction by CSVMs is based on the prediction method from the clustering algorithm. At first, the sequence segments whose structures to be predicted are assigned to a specific cluster in the cluster group by the clustering algorithm. The clustering algorithm and cluster membership assignment can be found in [25, 26]. Then CSVM modeled for this specific cluster is used to identify how close this sequence segment is nonlinearly correlated to the representative structure of this cluster. If the sequence segment is predicted as the positive sample by CSVM, this segment has the potential to be closely mapped to the representative structure for this cluster. Consequently, the representative local tertiary structure of this cluster can be safely assigned to this sequence segment. The method to decide the representative local tertiary structure of each cluster can be found in [25, 26]. If the sequence segment is predicted as the negative sample by CSVMs, this segment does not closely correspond to the local tertiary structure for this cluster. The structure of this segment cannot be reliably predicted. The assigned cluster will not be considered in the next iteration. In the next iteration, the cluster membership function is used to select the next cluster from the remaining clusters of the cluster group. The previous procedure will be repeated until one SVM modeled for the selected cluster predict the given sequence segment as positive and structure prediction is performed. The complete prediction process is shown in figure 1. CSVMs are used to reclassify sequence segments, which are misclassified by the conventional clustering algorithms.

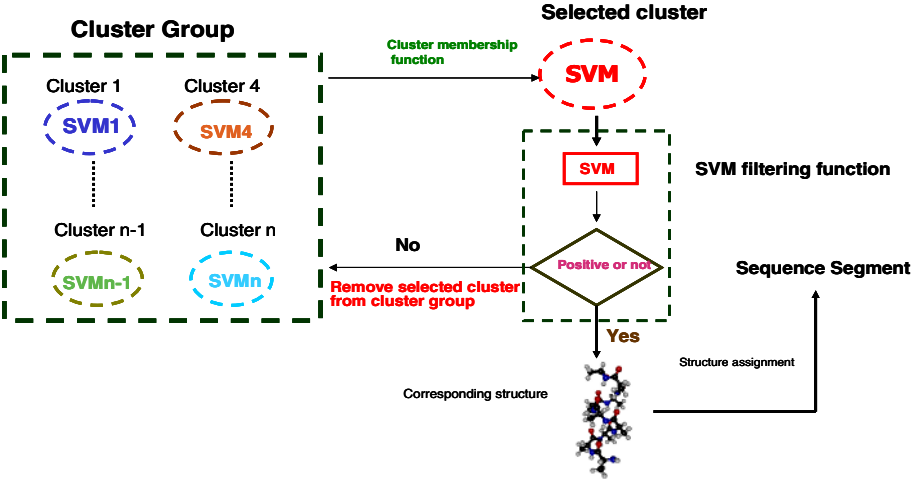

Fig. 1. Local Protein Structure Prediction by CSVMs Algorithm

3 Experimental Evaluation

3.1 Training Set and Independent Test Set

The training dataset used in our work includes 2090 protein sequences obtained from the Protein Sequence-Culling Server (PISCES)[20]. The training set is utilized to construct the CSVM model. 200 protein sequences from the recent release of PISCES are included into the independent test set. The test set is used in the performance evaluation phase. The structures of protein sequences in the training set and the testing set are available from Protein Data Bank (PDB) [4]. Any two sequences in the training set and the test set share less than 25% similarity.

3.2 Experimental Setup

Structure prediction accuracy for sequence segments in terms of secondary structure accuracy, Distance Matrix Root Mean Square Deviation (dmRMSD) and Torsion angle RMSD (taRMSD) is calculated to evaluate the performance of the conventional clustering algorithm and our new computational model. The definition for representative structures of clusters was introduced in [25, 26]. We use the formula in [11, 12, 24] to calculate secondary structure accuracy, dmRMSD, and taRMSD.

During the prediction process, structures of sequence segments are first predicted by clusters with the high training accuracy. If the structures of sequence segments cannot be predicted by clusters with high training accuracy, clusters with the lower training accuracy will be used for structure prediction.

Training secondary structure accuracy for a given cluster is the average training accuracy of sequence segments in the training set predicated by this cluster. Training dmRMSD of a given cluster is the average training dmRMSD of sequence segments in the training set predicated by this cluster. Training taRMSD of a given cluster is similarly defined. Test secondary structure accuracy, test dmRMSD and test taRMSD is similarly defined for each cluster in the independent test set.

Table 1 shows the standard to classify clusters into different groups based on the training accuracy of the clustering algorithm. In the good cluster group, all clusters have training secondary structure accuracy greater than 80%, training dmRMSD less than 1 Å and training taRMSD less than 25 degree. The bad cluster group and the average cluster group are similarly defined. As a result, the good cluster group includes all the clusters with highest training accuracy. The bad cluster group includes clusters with poor training accuracy.

Table 1. Standard to classify clusters into different groups

Cluster Group	Secondary Structure Accuracy	dmRMSD	taRMSD
Bad Cluster Group	between 60% and 70%	greater than 1.5 Å	greater than 30 degree
Average Cluster Group	between 70% and 80%	between 1 Å and 1.5 Å	between 25 and 30 degree
Good Cluster Group	greater than 80%	Less than 1 Å	less than 25 degree

As described in [25, 26], only combined information of secondary structure, torsion angle and distance matrix can represent protein structure precisely. In order to rigorously evaluate the prediction quality for these algorithms, we designed accuracy criterion for each cluster. Accuracy criteria for one cluster is the percentage of sequence segments with secondary structure accuracy greater than 70%, dmRMSD less than 1.5 Å and taRMSD less than 30 degree in the test set for this cluster. This accuracy criterion for one cluster reflects the percentage of sequence segments with the acceptable level of structure prediction accuracy.

3.3 Result and Analysis

By the experiment, we get the average accuracy, prediction and recall of SVMs for different cluster groups in Table 2. Average SVM accuracy for different cluster groups is over 80% in Table 2. This indicates that the generalization power for CSVMs is strong enough to recognize the complicated pattern of the sequence-to-structure relationship for each cluster. Table 3 compares the prediction accuracy between the conventional clustering algorithm and the CSVMs model using the definition for the accuracy criterion defined in the section 3.2.

Table 2. Comparison of the average accuracy, precision and recall of SVMs for different cluster groups

Cluster Group	Average Accuracy	Average Precision	Average Recall
Bad Cluster Group	75%	74%	77%
Average Cluster Group	82%	84%	95%
Good Cluster Group	85%	86%	96%

Table 3. Comparison of the accuracy between the conventional clustering algorithm and the CSVMs model for different cluster groups

Cluster Group	Conversional Clustering Algorithm	CSVMs Model
Bad Cluster Group	57.13%	62.34%
Average Cluster Group	68.76%	70.43%
Good Cluster Group	78.95%	79.34%

Table 3 shows that the prediction accuracy for local protein tertiary structure has improved by average 2% when the CSVMs model is applied. Since K-means clustering may introduce noisy and irreverent information into each cluster, the machine learning techniques are required to identify the strength of the sequence-to-structure relationship for each sequence segment. After learning the relationship between sequence distribution patterns and representative structures of each cluster, CSVMs can filter out potentially unreliable prediction and selects potentially reliable prediction for each cluster. Our experimental results indicate that building CSVMs for each granule respectively can increase effectiveness and efficiency of data mining algorithms.

4 Conclusion

SVM is not efficient for very large datasets due to the high training time complexity. To solve this problem, a new model called Clustering Support Vector Machines (CSVMs) is proposed. The special characteristics of CSVMs convert a complex classification problem into multiple smaller computational problems so that learning tasks for each CSVM are more specific and efficient. Each CSVM can concentrate on highly related samples in each cluster without being distracted by noisy data from other clusters. Because of data partitioning, the training tasks for each CSVM are parallelized. The parallel training process makes the data-mining task for very large datasets possible. The satisfactory experimental results for local protein structure prediction show that our computational model opens a new approach for solving the complex classification problem for huge datasets.

Further improvement for the CSVMs model will be made in the future work. Under the framework of granular computing, there are many granulation methods such as fuzzy sets, probabilistic sets, decision trees, clusters and association rules. In the future work, the more effective granulation method need be studied.

References

1. Agarwal, D. K.: Shrinkage estimator generalizations of proximal support vector machines. in Proc.of the 8th ACM SIGKDD international conference of knowledge Discovery and data mining, Edmonton, Canada (2002)
2. Award, M., Khan, L. Bastani, F. and Yen, I.: An Effective Support Vector Machines(SVMs)Performance Using Hierarchical Clustering. in Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2004).
3. Balcazar, J. L., Dai, Y. and Watanabe, O.: Provably Fast Training Algorithms for Support Vector Machines. in Proc.of the 1stIEEE International Conference on Data mining, IEEE Computer Society (2001) 43-50
4. Berman, H. M., Westbrook, J. and Bourne, P. E. :The protein data bank. Nucleic Acids Research, Vol. 28, (2000) 235-24
5. Bystroff, C. and Baker,D.:Prediction of local structure in proteins using a library of sequence-structure motifs. J. Mol. Biol, vol. 281, (1998) 565-77
6. Bystroff, C., Thorsson, V. and Baker, D.: HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. J. Mol. Biol, vol. 301, (2000) 173-90
7. Chang, C. C and Lin, C. J.: Training nu-support vector classifiers: Theory and algorithms. Neural Computations, vol. 13, (2001) 2119-2147
8. Cristianini, N., Shawe-Taylor,J.: An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK (2000)
9. Daniael, B. and Cao, D.: Training Support Vector Machines Using Adaptive Clustering. in Proc. of SIAM International Conference on Data Mining 2004,Lake Buena Vista, FL, USA.
10. Gupta, S. K., Rao, K. S. and Bhatnagar, V.:K-means clustering algorithm for categorical attributes. Data Warehousing and Knowledge Discovery DaWaK-99, Florence, Italy (1999) 203-208

11. Hu, H., Pan, Y., Harrsion, R. and Tai, P. C.: Improved protein secondary structure prediction using support vector machine with a new encoding scheme and advanced tertiary classifier. IEEE Transactions on NanoBioscience, Vol. 2, (2004) 265-271
12. Kolodny, R. and Linial, N.:Approximate protein structural alignment in polynomial time. Proc Natl. Acad. Sci., Vol. 101, (2004)12201-12206
13. Osuna, E. Freund, R. and Girosi, F.: An improved training algorithm for support vector machines. In Proc. Of IEEE Workshop on Neural Networks for Signal Processing,Pages (1997)276-285
14. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In advances in Kerenel Methods-Support Vector Learning, (1999) 185-208
15. Schoelkopf, B., Tsuda, K. and Vert, J. P.: Kernel Methods in Computational Biology. MIT Press, (2004)71-92
16. Scholkopf, B., Burges, C. and Smola, A. (eds); Advances in Kernel Methods-Support Vector Learning. MIT Press,Cambridge,MA(1999)
17. Valentini, G. and Dietterich,T. G.: Low Bias Bagged Support vector Machines. in Proc. of the 20th International Conference on Machine Learning ICML 2003,Washington D.C.USA,(2003)752-759
18. Vapnik, V. : Statistical Learning Theory. John Wiley&Sons, Inc., New York (1998)
19. Vavasis,S.A.: Nonlinear Optimization: Complexity Issues. New York: Oxford Science.(1991).
20. Wang, G. and Dunbrack, RL. Jr.:PISCES: a protein sequence-culling server. Bioinformatics, vol. 19, no. 12, (2003) 1589-1591
21. Yao,Y. Y: Granular Computing. Computer Science (Ji Suan Ji Ke Xue), Proceedings of The 4th Chinese National Conference on Rough Sets and Soft Computing, Vol. 31 (2004) 1-5
22. Yao,Y. Y.: Perspectives of Granular Computing. to appear in 2005 IEEE Conference on Granular Computing.
23. Yu, H., Yang, J. and Han, J.: Classifying Large Data sets Using SVMs with Hierarchical Clusters. in Proc. Of the 9th ACM SIGKDD 2003, Washington DC, USA (2003)
24. Zagrovic B. and Pande, V. S.:How does averaging affect protein structure comparison on the ensemble level? Biophysical Journal, Vol. 87, (2004) 2240-2246
25. Zhong, W., Altun, G., Harrison, R., Tai, P. C. and Pan, Y.: Mining Protein Sequence Motifs Representing Common 3D Structures. Poster Paper of IEEE Computational Systems Bioinformatics (CSB2005), Stanford University (2005)
26. Zhong, W., Altun, G., Harrison, R., Tai, P. C., and Pan, Y.: Improved K-means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property. IEEE Transactions on NanoBioscience, Vol. 4, (2005) 255-65