Volume 2, Issue 11, November 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Multiclass Classification for Protein Fold Prediction Using Smote

**K. Suvarnavani***, *Department of CSE,V.R.Siddhartha Engineering College*
**S. B. Rafiah**, *Department of CSE, P.V.P.Siddhartha Institute of Technology*
**Dr. Nageswara Rao .K** , *Department of CSE, P.V.P.Siddhartha Institute of Technology*

***Abstract— An approach to classify imbalanced data is described. Imbalanced classification is common problem found in fraud detection, rare disease diagnosis, network intrusion and others. A dataset is imbalanced if the classification categories are not approximately equally represented. Protein contact maps are two dimensional representations of protein structures. This paper addresses the problem of protein fold prediction which is a multiclass problem having imbalanced classes. There are two methods to solve this problem. One is considering all classes in one optimization and second is to cover one n-class problem into several binary class problems. Boosting algorithm, SMOTE is applied to rebalance the data set to boost the performance and then a decision tree classifier is used to classify "folds" from the features of contact map. This project shows that classification of unbalanced data has attained over 70% of accuracy for the feature set generated by Triangle Sub division method. TSM, a simple algorithm is implemented to extract additional features from contact map. The experimental results show that the classification performance of SMOTE is better even if all the 27 folds of four major structural classes of SCOP are included. It is also proved that generated feature set from TSM is well suited to carry out investigation since results obtained are substantial.***

***Keywords— Protein contact map, Imbalanced data, multiclass classification, boosting, SCOP.***

## I. INTRODUCTION

Multiclass classification problem with imbalanced data is a major problem in machine learning as applications that require discrimination among several classes are ubiquitous. Suppose we have a set of training data $(x_1,c_1),..(x_n,c_n)$ where the input (predictor variable) xi belongs to $R^p$ and the output (response variable) $C_i$ is qualitative and assumes values in a finite set. e.g.{1,2,….k} where k is no of classes. The goal is to find a classification rule C(x) from training data so that when given a new input x, we can assign it a class label c from {1,….K}.Many multi-class pattern classification techniques are proposed for learning from balanced datasets [4].Protein fold data set contains proteins from 27 most populated folds (classes) representing all major structural classes: all alpha, all beta, alpha + beta and alpha / beta. Protein fold prediction is a multi-class classification problem having highly imbalanced data amongst its classes. When more no of classes are involved it is more difficult to accurately predict the fold for a query sequence. The main aim of this paper is to find a solution for the multi-class protein fold prediction problem with the classes being highly imbalanced. Multi-class pattern classification techniques are principally proposed for learning from balanced dataset. Support Vector Machines are believed to be less prone to the class imbalance problem than other classification learning algorithms [6], since boundaries between classes are calculated with respect to only a few support vectors and the class sizes may not affect the class boundary too much. The reason for this phenomenon is that as the training dataset gets more imbalanced, the support vector ratio between the high sample class and the small sample class also becomes more imbalanced. Support Vector Machines simply learn to classify everything as the high sample class in order to make the margin the largest and the error the minimum.

## II. RELATED WORK

The dataset is said to be imbalanced dataset if the data samples are not equally distributed among classes. Kubat and Matwin (1997) proposed a technique in which the majority class was under sampled while the minority class was original. Here, geometric mean has been used as a performance measure [7]. Ling and Li (1998) combined over sampling of minority class with under sampling of majority class. Here, Lift analysis has been used to measure a classifier's performance [8]. Japkowicz (2000) evaluated three strategies for imbalanced data. The three techniques are under sampling, re-sampling and a recognition based induction scheme. Re-sampling was again classified as random re-sampling and focused re-sampling. Random re-sampling re samples the smaller class at random till its samples were as many as majority class. Focused re-sampling re-samples only those minority classes that are between majority and minority class [6].

## III. BOOSTING ALGORITHMS

Boosting is a admired technique for improving the accuracy of any given machine learning algorithm. Literature reports that AdaBoost and SMOTE algorithm have been successfully applied to most popular classifiers [ 9, 11, 12 ]. In literature, different boosting algorithms have been applied to unbalanced datasets consisting of a certain class of interest having very small size. Among the boosting algorithms AdaBoost [9] is shown to decrease the prediction error of minority classes. Significantly with increasing the prediction error of majority class a little bit, it can also produce higher values of margin which indicates a better classification. On the other hand, the combination of boosting and bagging, construct multiple classifiers by re sampling the data space, weighting samples by boosting and replacing samples by

bagging. The improvement in performance arising from ensemble combinations is usually the result of a reduction in variance. Variance measures how much a learning algorithm's guess bounces around for different training sets. Variance is therefore associated with over fitting, if a method over fits the data, the predictions for a single instance will vary between samples. Both boosting and bagging are capable of reducing variance, and hence are immune to the model over fitting problem. With an imbalanced data set, small class samples occurring infrequently, models that describe the rare classes have to be highly specialized. Standard learning methods pay less attention to the rare samples as they try to extract the regularities from the dataset. Such a model performs poorly on the rare class due to the introduced bias error. Bagging is believed to be effective for variance reduction, but not for bias reduction. AdaBoost and SMOTE, however are stated to be capable of both bias and variance reduction. There are many other boosting algorithms for classification of imbalanced data in the literature. They are categorized into two types.Type1 can be applied to classifier learning algorithms directly, such as AdaCost [9], CSB1 and CSB2 [10], and RareBoost [11].Type2 is combination of the data synthesis algorithm and the boosting procedures, such as SMOTEBoost [20], and DataBoostIM [14]. AdaBoost algorithm is an iterative procedure that combines many weak classifiers to approximate the bayes classifier C*(x).Starting with unweighted training sample, AdaBoost builds a classifier that produces class labels. If a training data point is misclassified , the weight of that training data point is increased(boosted).A second classifier is built using new weights which are no longer equal. Again misclassified training data have their weights boosted and the procedure is repeated. we may build any number of classifiers this way. A score is assigned to each classifier and the final classifier is defined as the linear combination of classifiers from each stage. We considered boosting algorithms that can be applied directly to most classification learning algorithms. In this category, Ada-Cost[9], CSB1 and CSB2 employ cost items to bias the boosting towards the small class, and RareBoost. Multiclass AdaBoost (SAMME): Stagewise additive modeling using a multiclass exponential loss function [21].This algorithm puts more weight on the misclassified data points and it also combines weak classifiers a little differently from AdaBoost. Here misclassified training samples get more weights. ABC Boost: It is based on two key ideas.

1. For multiclass classification, popular loss functions for k classes usually assume a constraint such that only the values for k-1 classes are needed. Therefore we can choose a base class and derive algorithms only for k-1 classes.
2. Though base class is not explicitly trained it is benefitted implicitly from training on k-1 classes. Thus base class is chosen adaptively.

## IV. SMOTE

Synthetic Minority Oversampling Technique is an oversampling approach in which the minority class is oversampled by creating synthetic examples rather than oversampling with replacement [12]. Extra training data can be created by performing certain operations. Depending upon the amount of over sampling required, neighbors from the k nearest neighbors are randomly chosen. Synthetic samples are generated as follows:

1. Take the difference between the feature vector (sample) under consideration and its nearest neighbor.
2. Multiply this difference by a random number between 0 and 1 and add it to the feature vector under consideration.

In SMOTE, Firstly, the minority class is oversampled by taking samples from the minority class and then introducing synthetic examples along the nearest neighbors of the minority class. Secondly, the majority class is under sampled by randomly removing samples from majority class until the minority class becomes some specified percentage of the majority class. Thus a combination of under-sampling and over-sampling leads to the initial bias of the learner towards the majority (negative) class being reversed in favor of the minority (positive) class. The classifiers are generally taken to be SVM and NN learned on the dataset by "SMOTING" the minority class and under-sampling the majority class [13].

### A. SMOTE Algorithm

```
%input: T minority sample; K amount of smote K%
%output: new_T synthetic sample
function new_T = smote(T, K)
[a, b] = size(T);
out_id = zeros(a, K);
dist = zeros(a, 1);
added_T = zeros(a, b, K);
for i = 1:a   % for each point
for j = 1:a
dist(j) = sqrt(sum((T(i,:)-T(j,:)).^2));
end
[y, index] = sort(dist);
out_id(i,:)=index(2:K+1)';
for jj = 1:K
alpha = rand;
added_T(i,:,jj)= alpha*T(i,:)+(1-alpha)*T(out_id(i,jj),:);
end
end
new_T = [];
```

```
for i = 1:K
new_T = cat(1, new_T, added_T(:,:,i));
end
```

## V. PROTEIN CONTACT MAP

The two dimensional representation of 3D structure is called a contact network which is a symmetric square matrix. The contact matrix C is obtained by taking a map from the amino acid (AA) sequence of the protein to itself with C(a, b) = 1 if amino acid residue 'a' is in proximity with 'b', say within 7 Angstrom units, otherwise zero. Protein contact maps are intermediary representations which are found to be useful to analyze the structural properties of proteins. Our work involves extracting useful features from the contact maps and utilizes them effectively for the protein fold prediction problem. Without loss of generality consider the lower triangular matrix for this study.

## VI. METHODOLOGY

Ding et al. considered features from structures and constructed a feature vector of dimension 20. Shamim et al. consider some additional features adding up upto 100 and also considered both sequence and secondary structural features [22]. More recently, Chmielnicki and et al. apply the multi-class support vector machine classifier for protein fold recognition problem [1]. Suvarna et al. applied boosting algorithm to the feature set generated by eight neighboring algorithm [3]. In this paper, SMOTE is applied to the feature set generated by TSM algorithm.

## VII. DATASET

The data set of Ding et al is considered as a bench mark data in the literature [15], for protein folds. This data set contains proteins with less than 35-40% sequence identity representing all classes and further contains proteins belonging to 27 most populated folds, representing all major structural classes of all alpha, all beta, alpha +beta and alpha/beta. Along with the data set of Ding et al the recently updated database is used that holds consensus view of fold space based on SCOP [16], CATH and DALI [5, 20]. This database is designed such that proteins domains are classified as a protein fold if they agree in at least two of three classification systems. These proteins are then downloaded from the protein data bank [18, 19] for which the contact map is computed. The contact matrices of the map form the input data set for the classification experiments. Useful features are to be computed for these proteins. In the literature, amino acid composition, physico-chemical properties and secondary structural properties have been considered as features for classification. Ding [15] et al. Consider 20 dimension feature vectors. We extracted novel contact map features [3] along with the features proposed to form twenty dimensional feature vector representations. We choose testing set to have 244 sequences and training set with 296 sequences.

## VIII. FEATURE EXTRACTION

A feature extraction scheme, called TSM [2] is implemented, which helps in identifying clusters. The contact network is a symmetric square matrix. The lower tri-angular matrix is considered for our experimentation. The very first step is to divide the triangle into four equal parts and assign a label L for left, M for middle, R for right, and T for top triangles. These patterns are helices and sheets. Here, we extracted the dense pattern regions. So, the sub patterns are Helix-Helix, Helix-Sheet, Sheet-Sheet, and Sheet-Coil. The protein contact map can be represented using density of the contacts in the sub triangles. This method is independent of the length of the protein and can be used to build a feature vector. In order to discriminate effectively, we need to iterate this procedure to one more level and subdivide each triangle into four sub-triangles and annotate the clusters within R with labels RL, RM, RR, and RT. Similarly, the triangles within L, M, and T are labeled with LL, LM, LR, LT; ML, MM, MR, MT; and TL, TM, TR, TT, respectively. Each protein contact map is represented as a feature vector of length of 20. In this project, 27 most populated folds in the database which have minimum of seven or more proteins. They represent all major structural classes all alpha, all beta, alpha/beta, alpha+beta. All the instances of four major structural classes were increased to 1485.

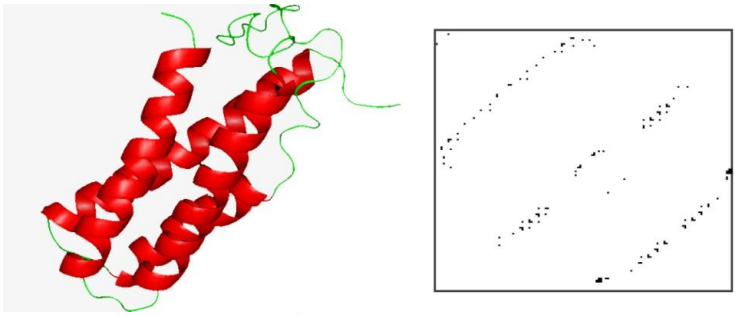


Figure 1: 3D Structure of 1BBH protein and its masked contact map.

The 3D structures and the masked contact maps of proteins with protein_id: 1BBH, 3HHR, 1GCA, 1FXR of All alpha, All Beta, Alpha /Beta and Alpha + Beta respectively are included. The 1BBH protein belongs to all alpha class whose fold is 4 helical up and down bundle. The contact maps of four proteins of four different structural classes respectively are generated and the information present along the diagonal is masked since the idea is to extract the information from off diagonal contact map.

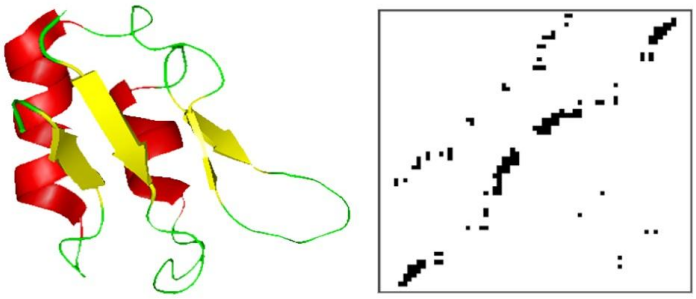


Figure 2: 3D Structure of 1FXR protein and its masked contact map.

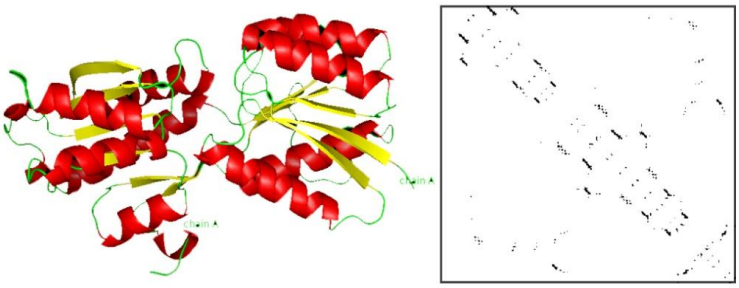


Figure 3: 3D Structure of 1GCA protein and its masked contact map.

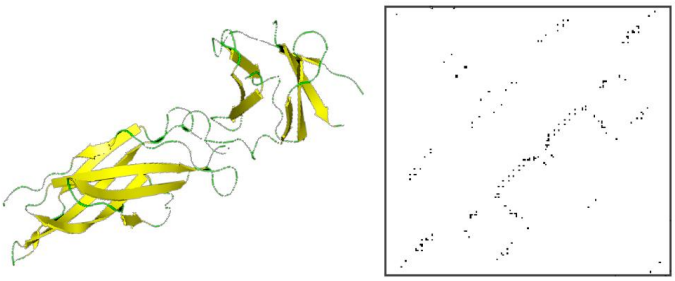


Figure 4: 3D Structure of 1FXR protein and its masked contact map.

## IX. CLASSIFICATION RESULTS

We have done classification without using SMOTE and found that accuracy obtained is very low. The classification is done using J48 decision tree which is available in open source software WEKA [17]. The highlight of this study is generated feature set has given good accuracies and hence it can be concluded that further investigation can be carried out by using feature set of TSM. Table I shows that the accuracy of classification has been increased when SMOTE is applied to this extracted feature set of TSM. SMOTE is an algorithm which rebalances the inadequate classes. Initially it takes the maximum instance class in the training set and copes up the remaining class samples into that ratio maintained by the smote parameter 'K'. All the four structural classes were included and the number of instances of all the classes together was 1485. A decision tree classifier J48 learning algorithm which is in WEKA is used for classification. Table II gives results of classification of All Alpha structural class. In the similar way Table III, Table IV, Table V shows the results of evaluation measure for All beta, Alpha/Beta and Alpha + beta respectively.Table I shows that shamim et al. has got an accuracy of 53.92%. The size of feature vector of shamim et al. is 125.Ding et al. attained accuracy of 40.46% with feature vector of size 21.Similarly Suvarna et al. has attained an accuracy of 65.34% with feature vector of size 11.It is shown in the table that 78.25% of accuracy has obtained when TSM feature set of size 20 is used.

TABLE I: CLASSIFICATION OF SCOP STRUCTURAL CLASSIFICATION OF A PROTEIN INTO 4 MAJOR STRUCTURAL CLASSES.

| STRUCTURAL CLASS | SHAMIM ET AL.[22] | DING ET AL. | SUVARNA ET AL. | SMOTE WITH TSM FEATURE SET |
|---|---|---|---|---|
| ALL ALPHA | 78 | 55 | 75.06 | 80 |
| ALL BETA | 56 | 33.18 | 69 | 78 |
| ALPHA+BETA | 35.8 | 31.4 | 63.3 | 84 |
| ALPHA/BETA | 45.9 | 42.26 | 54 | 71 |
| AVERAGE | 53.925 | 40.46 | 65.34 | **78.25** |

TABLE III: CLASSIFICATION OF ALL ALPHA CLASS.

| TP | FP | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.833 | 0.044 | 0.789 | 0.833 | 0.811 | 0.897 | 1 |
| 0.821 | 0.033 | 0.836 | 0.821 | 0.829 | 0.907 | 3 |
| 0.75 | 0.07 | 0.689 | 0.75 | 0.718 | 0.865 | 4 |
| 0.883 | 0.03 | 0.869 | 0.883 | 0.876 | 0.95 | 7 |
| 0.771 | 0.036 | 0.787 | 0.771 | 0.779 | 0.871 | 9 |
| 0.731 | 0.026 | 0.844 | 0.731 | 0.784 | 0.891 | 11 |

TABLE III: CLASSIFICATION OF ALL BETA

| TP | FP | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.623 | 0.04 | 0.66 | 0.623 | 0.641 | 0.813 | 20 |
| 0.764 | 0.016 | 0.857 | 0.764 | 0.808 | 0.91 | 23 |
| 0.813 | 0.028 | 0.765 | 0.813 | 0.788 | 0.906 | 26 |
| 0.782 | 0.023 | 0.811 | 0.782 | 0.796 | 0.896 | 30 |
| 0.962 | 0.014 | 0.893 | 0.962 | 0.926 | 0.976 | 31 |
| 0.6 | 0.053 | 0.566 | 0.6 | 0.583 | 0.848 | 32 |
| 0.909 | 0.021 | 0.847 | 0.909 | 0.877 | 0.955 | 33 |
| 0.893 | 0.019 | 0.862 | 0.893 | 0.877 | 0.948 | 35 |
| 0.729 | 0.026 | 0.796 | 0.729 | 0.761 | 0.87 | 39 |

TABLE IV: CLASSIFICATION OF ALPHA / BETA

| TP | FP | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.644 | 0.058 | 0.594 | 0.644 | 0.618 | 0.826 | 46 |
| 0.704 | 0.033 | 0.717 | 0.704 | 0.71 | 0.895 | 48 |
| 0.672 | 0.061 | 0.614 | 0.672 | 0.642 | 0.838 | 51 |
| 0.789 | 0.038 | 0.726 | 0.789 | 0.756 | 0.913 | 54 |
| 0.75 | 0.031 | 0.763 | 0.75 | 0.756 | 0.869 | 57 |
| 0.563 | 0.035 | 0.628 | 0.563 | 0.593 | 0.778 | 59 |
| 0.764 | 0.02 | 0.824 | 0.764 | 0.792 | 0.886 | 62 |
| 0.911 | 0.013 | 0.895 | 0.911 | 0.903 | 0.948 | 69 |

TABLE V: CLASSIFICATION OF ALPHA+BETA

| TP | FP | PRECISION | RECALL | F-MEASURE | ROC AREA | CLASS |
|---|---|---|---|---|---|---|
| 0.88 | 0.067 | 0.846 | 0.88 | 0.863 | 0.917 | 72 |
| 0.845 | 0.071 | 0.86 | 0.845 | 0.852 | 0.925 | 87 |
| 0.806 | 0.102 | 0.82 | 0.806 | 0.813 | 0.892 | 110 |

## X. EVALUATION MEASURES FOR PERFORMANCE PREDICTION

A confusion matrix contains information about the actual and predicted classifications done by a classification machine learning algorithm. Performance of algorithm is evaluated using the data available in the confusion matrix. True Positive Rate (TP) is the proportion of positive cases that were correctly predicted. False positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive. The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly. False negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative. The predictions are evaluated by using the fivefold cross-validation and carried out to

show the robustness of the results. Recall (True Positive Rate) gives the proportion of positives out of the total positions predicted as positive and is calculated as

$$\text{Recall} = \frac{TP}{TP+FN}$$

Precision is the percentage of positive predictions that are correct. It is calculated as

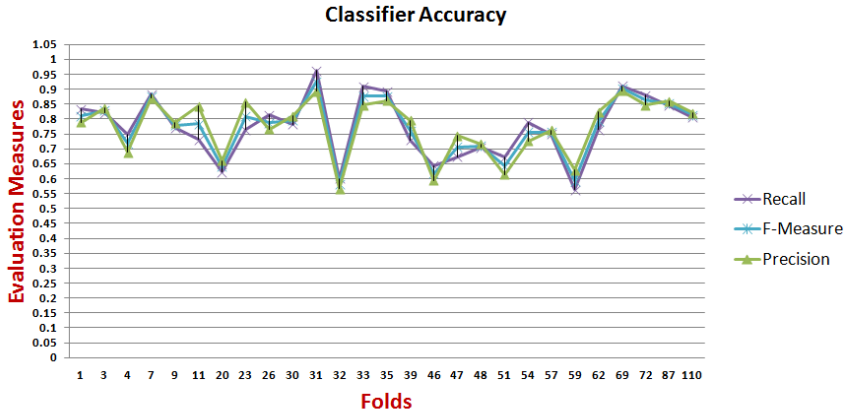
$$\text{Precision} = \frac{TP}{TP+FP}$$

F-Measure is calculated using following formula

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The graph which is shown below shows the performance of classification. The evaluation measures considered are Precision, Recall and F-Measure which are obtained from confusion matrix. The values of these parameters vary from 0.5 to 1.

FIGURE 5:GRAPH WHICH SHOWS PERFORMANCE OF CLASSIFIER ACCURACY




## XI. CONCLUSION

In this project, several important issues related to protein fold prediction are studied. Literature shows that multiclass classification for unbalanced data is ongoing research problem. Various multiclass classification techniques like one versus one, one versus all, all versus all etc exist. Classification cannot be obtained accurately when numbers of classes are more. Binary classification should be extended to classify among multiple classes. While classifying multiple classes, if any of the classes with less number of instances exist, then the classification is not obtained accurately. So, to equally distribute the instances among classes, boosting technique is used. Here SMOTE is applied to rebalance the data. TSM is used to extract the feature set. The extracted feature set is best suited to carry out the investigation since results obtained are promising. SMOTE when used with TSM feature set has obtained classification accuracy of 78.25%.

## REFERENCES

1. W. Chmeilnicki,K. Stapor. An efficient multi-class support vector machine classifier for protein fold recognition, IWPACBB - 2010, 77-84.
2. S. D. Bhavani and K.Suvarnavani, Somdatta Sinha. Mining of protein contact maps for proteinfold prediction. WIREs Data Mining and Knowledge Discovery, John Wiley & Sons, Volume1, Pages 362-368, July/August 2011
3. Suvarnavani K and S. D. Bhavani, SMOTE based Protein fold prediction classification, Advances in Computing and Information Technology 2013.
4. Ghanem A.S., Venkatesh, S. West, G.: Multi-class Pattern Classification in Imbalanced Data.
5. Day R, Beck DAC, Armen RS, Daggett V. A consensus view of fold space: Combining SCOP, CATH, and the Dali domain dictionary, Protein Science 2003, 12:21502160.
6. N.,Japkowicz and S. Stephen. The class imbalance problem: A systematic study. Intelligent Data Analysis Journal, 6(5):429–450, November 2002.

7. M. Kubat and S. Matwin, ``Addressing the Curse of Imbalanced Training Sets: One Sided Selection," in Proceedings of the Fourteenth International Conference on Machine Learning, (Nashville, Tennesse), pp. 179-186, Morgan Kaufmann, 1997.
8. C. Ling and C. Li, ``Data Mining for Direct Marketing Problems and Solutions," in Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining *(KDD-98)*, (New York, NY), AAAI Press, 1998
9. W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan. Adacost: misclassification cost-sensitive boosting. In Proceedings of Sixth International Conference on Machine Learning(ICML-99), pages 97–105, Bled, Slovenia, 1999.
10. K. M. Ting. A comparative study of cost-sensitive boosting algorithms. In Proceedings of the 17th International Conference on Machine Learning, pages 983–990, Stanford University, CA, 2000.
11. M. V. Joshi, V. Kumar, and R. C. Agarwal. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In Proceeding of the First IEEE International Conference on Data Mining(ICDM'01), 2001.
12. Chawla, N.V., Bowyer, K.W., Hall L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique Journal of Arti_cial Intelligence Research 16 (2002) 321-357.
13. N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. SMOTEBoost: Improving prediction of the minority class in boosting. In Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Database, pages 107–119, Dubrovnik,Croatia, 2003.
14. H. Guo and H. L. Viktor. Learning from imbalanced data sets with boosting and data generation: The databoost-IM approach. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets, 6(1):30–39, 2004.
15. Ding C H Q, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics , 2001, 17:349-358.
16. Shi J-Y, Zhang Y-N. Fast SCOP classification of structural class and fold using secondary structure mining in distance matrix, PRIB2009, LNBI 2009, 5780:344-353.
17. http://www.cs.waikato.ac.nz/ml/weka/
18. http://www.dynameomics.org/
19. http://www.rcsb.org/pdb/home/home.do
20. http://scop.berkeley.edu/. (Accessed June 2010).
21. J. Zhu, H. Zou, S. Rosset, and T. Hastie, " Multiclass adaboost, " Int. J. of Statistics and Its Interface, vol. 2, pp. 349-360, 2009.
22. Shamim MTA, Anwaruddin M, Nagarajaram HA. Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs, Bioinformatics 2007, 23:24 3320-3327.