International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 2 (2017), pp. 213-224

http://www.ripublication.com

Visualize Biological Database for Protein in Homosapiens Using Classification Searching Models

N. Deepak Kumar
Dept. Of CSE, SV University, Tirupati, India.

Dr. A. Ramamohan Reddy
Dept. Of CSE, SV University, Tirupati, India,

Abstract

Visualize biological database for protein is very complicated without Classify the protein properties. Protein classification is one of the major application of machine learning algorithms in the field of bio-informatics. The searching classification model works in two steps. Firstly, the correlation based feature selection for protein classification will be taken and strongly correlated features will be considered for classification using MST based. In second step, using Robust Regression, the classification will be performed. Based on results of RRC algorithm, it is highly has classification ratio than traditional machine learning algorithms such as SVM, Naïve-bayes, Decision Trees.

Keywords: Regression, Protein, correlation, Bioinformatics, DBMS.

1. INTRODUCTION

Protein-classification is an important research content of biological information, which is currently a central issue in proteomics research. It usually occurs in the formation of complexes, and has a wide range of application in protein function prediction, drug design and disease diagnosis [1]. In present ,several primary properties of protein that characterize a protein-protein interaction, and these properties always contain irrelevant and redundant information, which may make

knowledge discovery during training more difficult, and seriously reduce the speed of classification and the recognition accuracy. Consequently, it is necessary to condense the dimension of substantial biological information from hundreds to fitting size, feature selection is a obligatory measure to take [2].In general, most learning algorithms require a vectorial representation of the data in terms of features. Representing the data through low-dimensional informative feature sets is critical for the accuracy and complexity of the algorithms. However, for many biological problems it is not yet under-stood which features are informative[3].

Towards this project, we introduce a new correlation based technique for feature selection. All the features will be tested each other. Based on the results, the strongly correlated features will be taken for further classification.

2. BACKGROUND

A. Feature Extraction for Protein Data

Extracting features from protein is an indispensable step in the protein classification [9]. One of the key objectives of extracting features from protein is to provide a method for converting non-numerical attributes in sequences to numerical attributes [10]. There are three main methods for coding a whole sequence. These three methods include the composition, profile and pairwise homology alignment methods [11].

1. The Composition Method : The composition method has been the most popular method for analysing whole protein sequences for many years. A protein sequence can be expressed by using a vector of 20 numerical attributes with composition method.
2. Profile Method : A protein sequence is expressed as a set of similarities (probabilities) with a model or a family of sequences by using profile method. There are two ways to generate profiles.
3. Pairwise Homology Method : In order to get the best classification effect, both positive and negative sequences should be used for learning. The homology alignment method has been the dominant method for SVM.

B. Classification Method for Imbalanced Data

Classification with imbalanced data presently represents a great challenge in machine learning domain . In the classification problem field, data sets are imbalanced when the numbers of examples that represent the different classes are very different.

A large number of approaches have been put forward to deal with the class imbalance problems. These work of addressing the class imbalance problems can be divided into two categories.

1. Data Level approach rebalance the distribution by re-sampling the data space. Various sampling techniques are used to create an artificially balanced distribution of class examples for training.
2. Algorithm Level approaches try to adapt specific classification algorithm to reinforce the learning towards the minority class.

3. BASIC CONCEPTS

3.A. Correlation

Correlation is widely used in machine learning and statistics for relevance analysis[4]. For a pair of features (X, Y), the linear correlation coefficient ρ is given by

$$\rho = \frac{\sum_i (x_{i-} \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \overline{x_i})^2} \sqrt{\sum_i (y_i - \overline{y_i})^2}}$$

Where

$$x_i = \frac{x}{n}, y_i = \frac{y}{n},$$

ρ value will be $-1 \leq \rho \leq 1$.

If X and Y are completely correlated, ρ takes the value of 1 or -1; if X and Y are independent, ρ is zero. It is a symmetrical measure for two features.

Based on the results of correlation method, the strongly correlated techniques will be selected. Based on the results we got fit size of reduced features will be used for protein classification.

3.B. Robust Regression

Regression methods is one of the most extensively used methods to cope with data analysis.

The Linear Regression Model formula as follows:

$$y = x_{i1}\beta_1^0 + x_{i2}\beta_2^0 + \cdots + x_{i\rho}\beta_\rho^0 - e^i$$

Robust regression can be used as a alternative for least squares regression. In Least Square regression, getting the large residual is one of the issue. Robust and resistant regression analyses provide options to a least squares model when the data defy the fundamental assumptions. Robust and resistant regression measures dampen the

influence of outliers, as compared to regular least squares estimation, in an effort to provide a better fit for the majority of data.

For n data $y_i = X_i^T \beta + \varepsilon_i$

$$\varepsilon_i(\beta) = y_i - X_i^T \beta$$

where $i=1,\ldots,n$. Here we have rewritten the error term as

$\epsilon i(\beta)$ = error term's dependency

the Least absolute Deviation Estimator(LAD) is

$$\hat{\beta}_{LAD} = arg\ min_{\beta} \sum_{i=1}^{n} |\varepsilon_{i(\beta)}|$$

Which minimizes the absolute value of the residuals (i.e., $|r_i|$)

Some basic annotations' are listed here.

[1]Residual = Predicted Value – Actual Value

[2]Outlier = observation with larger residual

[3]Leverage: Extreme value on Predictor variable.

[4]Influence: a observation which substantially affects the estimate of the regression coefficients.

[5]Cook's distance: A measure that combines the information of leverage and residual of the observation.

The ordinary least squares estimation for LR are optimal when all of the regression assumptions are valid. If some of these assumptions are invalid, LSR method can perform poorly. **Robust regression** methods provide an option to least squares regression by necessitating less restrictive estimations . These methods try to reduce the influence of outlying cases one by one to provide a better fit to the majority of the data.

4. ROBUST REGRESSION BASED CLASSIFICATION (RRC) ALGORITHM

When analyzing the relationship between features, we have a conjecture that perhaps the spatial structure of protein will affect the relationship between features. The strongly correlated features will be taken into classification for efficiency improvising.

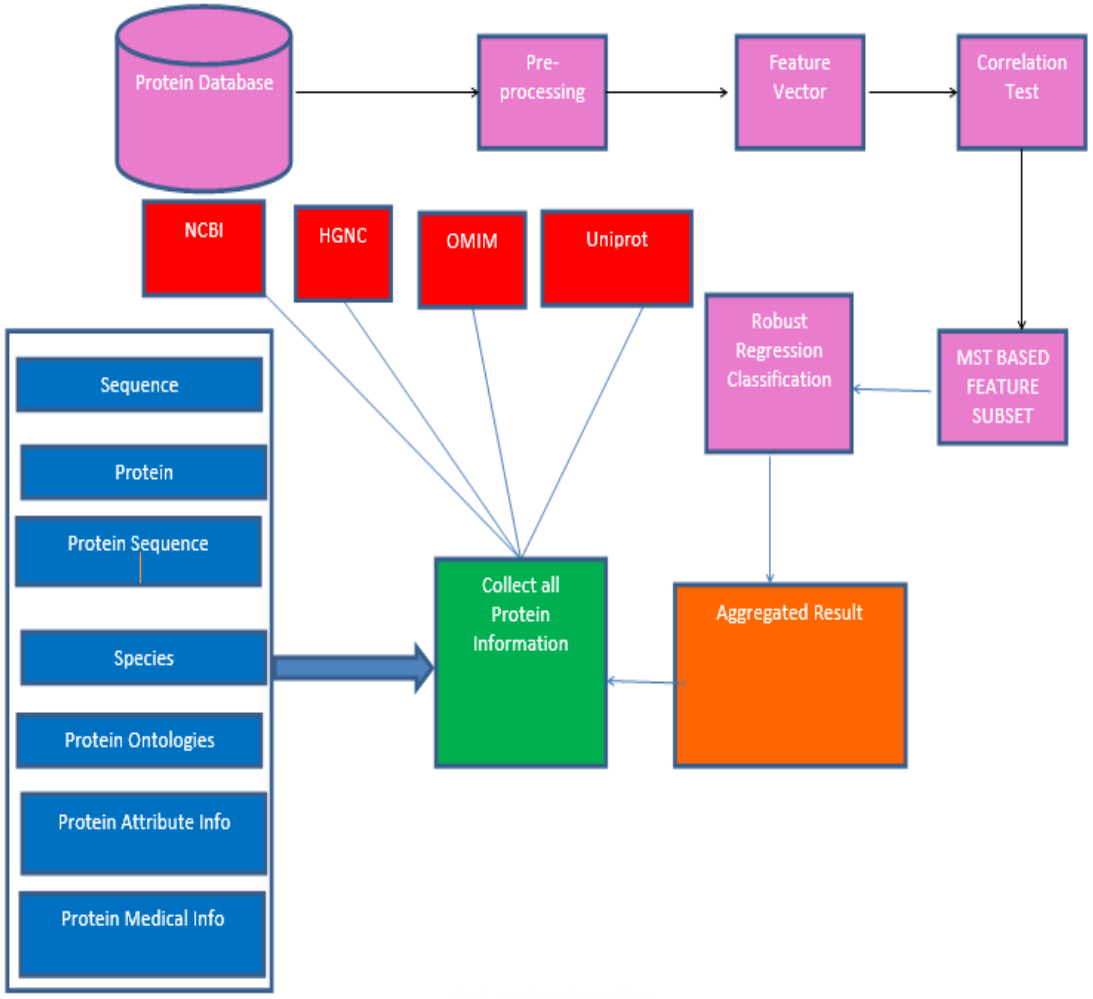

Fig 1. Architecture Flow

Protein Database:

Protein databases consisting of various protein information data (sequence id, cytogenetic location, protein name, protein id, structural id etc.) forms an essential component of this classification system. This data can be from single source or multiple sources. Protein database can have both relevant and irrelevant data from the view point of classification problem in hand. This has to be properly dealt by pre-processing.

Pre-Processing:

Data extracted from protein database needs to be pre-processed in order to make it suitable for further analytical processing. As huge no of records will be available and also the number of attributes associated with each record could also be high. Appropriate relevance analysis will help in choosing the records and attributes for

further analysis. Data pre-processing will deal with missing values, normalization and make the data suitable for further analysis.

Feature Vector:

Extracting features from protein is an indispensable step in the protein classification. One of the key objectives of extracting features from protein is to provide a method for converting non-numerical attributes in sequences to numerical attributes. The techniques used here are composition method for balanced data and data level approach for imbalanced data.

Correlation Test:

The feature vectors will be grouped as set named as S. Each feature will be tested with other features using correlation test. The correlation values will be calculated and monitored. All correlation values will be calculated and formed as correlation matrix. Strongly correlated features will be taken into consideration for further steps.

Feature Subset:

The strongly correlated features which are having values $0.5 \leq p \leq 0.5$ will be formed as a subset. The subset will consist of important features will be used in next steps for classification of proteins. The correlation between any pair of features and the target variable is referred to as the F-correlation. Thus for graph G, we build an MST, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well known Prim's algorithm.

Minimum Spanning Tree

1) Create a forest F (a set of trees), where each vertex in the graph is a separate tree.

2) Create a set S containing all the edges in the graph

3) While S is nonempty and F is not yet spanning

   a) remove an edge with minimum weight from S

   b) if that edge connects two different trees, then add it to the forest, combining two trees into a single tree

   c) otherwise discard that edge.

At the termination of the algorithm, the forest forms a minimum spanning forest of the graph. If the graph is connected, the forest has a single component and forms a minimum spanning tree. The sample tree is as follows, In this tree, the vertices represent the relevance value and the edges represent the F-Correlation value. The

complete graph G reflects the correlations among all the target-relevant features. Unfortunately, graph G has k vertices and k(k-1)/2 edges. For high-dimensional data, it is heavily dense and the edges with different weights are strongly interwoven. Moreover, the decomposition of complete graph is NP-hard. Thus for graph G, we build an MST, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well known Kruskal algorithm. The weight of edge (Fi`,Fj`) is F-Correlation SU(Fi`,Fj`).

After building the MST, in the third step, we first remove the edges whose weights are smaller than both of the T-Relevance SU(Fi`, C) and SU(Fj`, C), from the MST. After removing all the unnecessary edges, a forest Forest is obtained. Each tree Tj € Forest represents a cluster that is denoted as V (Tj), which is the vertex set of Tj as well. As illustrated above, the features in each cluster are redundant, so for each cluster V (Tj) we choose a representative feature Fj R whose T-Relevance SU(Fj R,C) is the greatest

1. G=NULL
2. F-correlation = (fi,fj)
3. Add fi,fj to G with F-correlation as weight of the corresponding node
4. Min span tree= Prim(G)

Robust Regression Algorithm

Robust Regression will be performed with S'. Perform least Square Estimation(LSE) to derive the leverage, residuals and outliers. Perform Iteratively Reweighted Least Squares(IRLS) to derive the influence and protein family group. Group the proteins based on features and influence the outliers.

Steps:

1. S'=NULL
2. Perform correlation with each one-to-one features and create a correlation matrix
3. Select the most strongly correlated features as a subset from S to form S'
4. According to the sequence order, perform Robust Regression in S' identify the protein family.
5. Perform least Square Estimation(LSE) to derive the leverage, residuals and outliers.
6. Perform Iteratively Reweighted Least Squares(IRLS) to derive the influence and protein family group.
7. Group the proteins based on features and influence the outliers
8. End.

Aggregated Results:

The aggregated results of RRC will be shown. Results will be compared with the results of SVM, Naïve bayes, Random Forest inorder to show the efficiency of RRC algorithm.

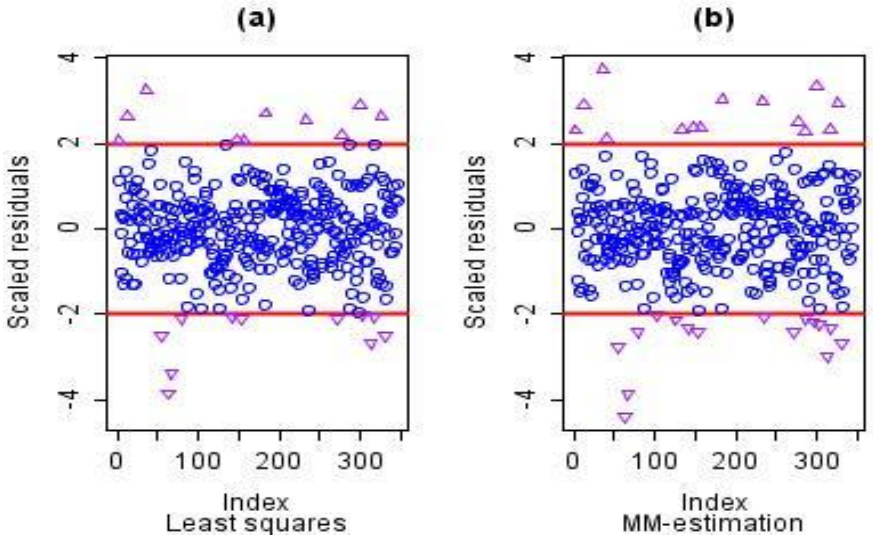

Fig 2. Robust Regression with Classification results.

The dark blue points represents the grouped and identified proteins family. The triangle points describes the outlier detection in the protein sequence feature data.

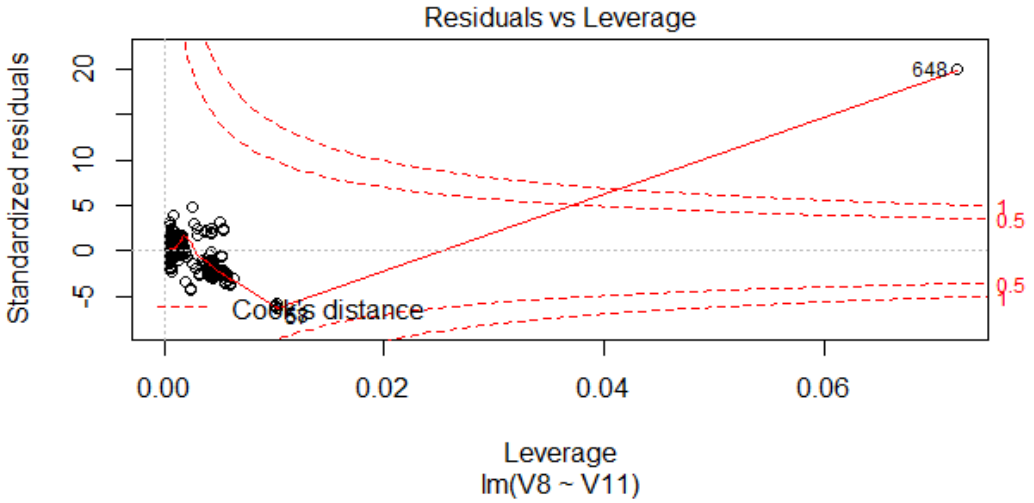

Fig 3. Cook distance measure for protein sequence.

The cook distance measure indicates that regression value is close well and outlier's are mentioned in the top points on the graph.

5. EMPIRICAL STUDY

We propose a novel method to classify the imbalanced protein data sets. The novel method constitutes a three-stage framework which consists of a feature extraction stage to rep-resent protein sequence, a method addressing class imbalance problem, and a method addressing class imbalance with good generalization ability.

To evaluate the usefulness of our RRC approach, we performed experiment on a training dataset that include 7839 samples and each sample has 51-dimension ASA features and 51-dimension SCORE features data. We calculate the correlation between the center feature and all other features for ASA and SCORE biochemical characteristics respectively. According to the calculation result, we built four subsets of features, whose size is 86, 76, 40 and 28 individually. We directly classified them using RRC algorithm in order to evaluate the Cooks distance and predictive accuracy.

We empirically evaluate the competence and efficiency of our RRC method by comparing it with traditional algorithms such as SVM, Naïve bayes, Neural Network and Random Forest. The following two figures show the result of comparisons between RRC and traditional algorithms.

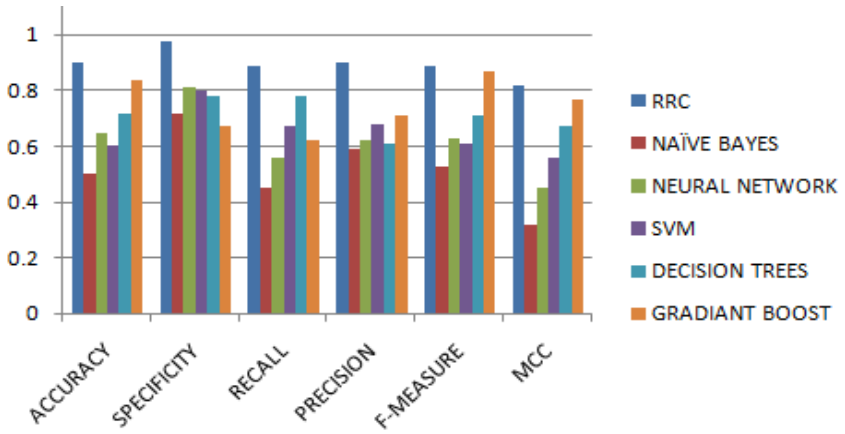

Fig 4. Evaluation of Performance Measure Matrices

Experimental evidence shows that RRC method has a better efficiency and effectiveness, especially in the regard of ROC Area. As shown in Fig. 2, while the feature dimension is reducted continuously, information in data is definitely significantly destroyed, this lead to ROC area in both the two methods dropped. Nevertheless, it is easy to see that the trend of ROC's decreasing in RRC method is

much more gentle than in Random Forest method. On the other hand, Fig. 2 shows that RRC demonstrates very high results with Naïve-bayes in the aspect of accuracy.

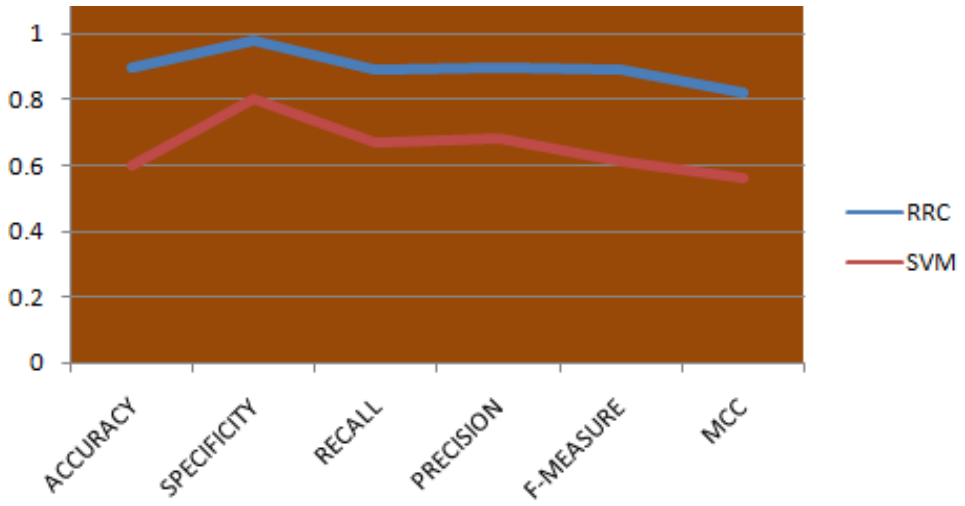

Fig 7. Comparison between RRC and SVM

The proposed technique also has shown improvements in the execution time by minimizing the overall classification time from few minutes to seconds. Moreover, Figure 2 shows the comparison and analysis of performance measure values using the proposed techniques. The Figure indicates that a reasonable improvement has been observed in the classification accuracy, specificity, Precision and F-Measure The improvements in the classification performance also results in minimizing the false alarm rate. The classification results indicates that the proposed technique would be helpful in the classification of distantly related or remotely homologus protein sequences which got ultimate interest of the research community.

6. CONCLUSION

The proposed Research work also has shown improvements in the execution time by minimizing the overall classification time from few minutes to seconds. A reasonable improvement has been observed in the classification accuracy, specificity, Precision and F-Measure the improvements in the classification performance also results in minimizing the false alarm rate. The classification results indicates that the proposed technique would be helpful in the classification of distantly related or remotely homologous protein sequences which got ultimate interest of the research community. After classification process is over and send related protein functionalities to related data base and easily visualize the protein information related biological databases(NCBI,UNIPROT,HGNC,OMIM).

REFERENCES

[1] C Ding, H Peng. "Minimum Redundancy Feature Selection from Microarray Gene Expression Data".Journal of Bioinformatics and Computational Biology, Vol. 3(2): pp.185–205, 2005.

[2] M. Robnik-Sikonja and I. Kononenko. "Theoretical and empirical analysis of Relief and Relief F." Machine Learning, 53:23–69, 2003.

[3] Y. Kim, W. Street, and F. Menczer. "Feature selection for unsupervised learning via evolutionary search". In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 365–369, 2000.

[4] A. Miller. "Subset Selection in Regression". Chapman & Hall/CRC, 2 edition, 2002.

[5] H. Almuallim and T. G. Dietterich. "Learning boolean concepts in the presence of many irrelevant features". Artificial Intelligence, 69(1- 2):279–305, 1994.

[6] M. Dash, K. Choi, P. Scheuermann, and H. Liu. "Feature selection for clustering – a filter solution". In Proceedings of the Second International Conference on Data Mining, pp. 115–122, 2002.

[7] P. Mitra, C. A. Murthy, and S. K. Pal. "Unsupervised feature selection using feature similarity". IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):301–312, 2002.

[8] H. Liu, H. Motoda, and L. Yu. "Feature selection with selective sampling". In Proceedings of the Nineteenth International Conference on Machine Learning, pp. 395–402, 2002.

[9] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, pp. 3446–3453, 2012.

[10] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *Blagus and Lusa BMC Bioinformatics*, vol. 14, p. 106, 2013.

[11] C. L. Castro and A. P. Braga, "Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data," *IEEE T. Neural Netw. Learn. Sys.*, vol. 24, pp. 888–899, 2013.

[12] W.-J. Lin and J. J. Chen, "Class-imbalanced classifiers for highdimensional data," *Brief. Bioinformatics*, vol. 14, pp. 13–26, 2012.

[13] S. Wang and X. Yao, "Multiclass imbalance problems: analysis and potential solutions," *IEEE T. Sys., Man, Cy. B*, vol. 42, pp. 1119–1130, 2012.

[14] H.-L. Dai, "The fuzzy Laplacian classifier," *Neurocomputing*, vol. 111, pp. 43–53, 2013.

[15] L. Gonzalez-Abrila, H. Nu˜nezb, C. Anguloc, and F. Velascoaa, "GSVM:An SVM for handling imbalanced accuracy between classes inbiclassification problems," *Appl. Soft Comput.*, vol. 17, pp. 23–31, 2014.

[16] M. A. Tahir, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification,"*Pattern Recogn.*, vol. 45, pp. 3738–3750, 2012.

[17] Z. R. Yang, "Biological applications of support vector machine," *Brief. Bioinform.*, vol. 5, pp. 328–338, 2004.