



Improved Classification Accuracy of Cancer Gene selection using Random Forest Tree and Neural Network

Shweta Parmar

M.TECH Scholar

Department of Computer Science & Engg.
PCST, Bhopal, India

Rajendra Patel

Asst. Prof

Department of Computer Science & Engg.
PCST, Bhopal, India

Abstract— In this paper we combined neural network with random forest ensemble classifier for classification of cancer gene selection for diagnose analysis of cancer diseases. The proposed method is different from most of the methods of ensemble classifier, which follow an input output paradigm of neural network, where the members of the ensemble are selected from a set of neural network classifier. the number of classifiers is determined during the rising procedure of the forest. Furthermore, the proposed method produces an ensemble not only correct, but also assorted, ensuring the two important properties that should characterize an ensemble classifier. In this paper we modified the sampling technique for cancer data classification using RF classification. For the modification of multiclass classification binary support vector classifier used. For the experimental process we used reputed dataset such dataset provided by UCI machine learning repository. Our proposed method implement in matlab 7.8.0. Our empirical result evaluation shows that better performance of our proposed method in compression of RF and DRF.

Keywords— ensemble Classification, Decision tree, , Random forest, Machine learning

I. INTRODUCTION

Normally the number of genes (features) is much greater than the number of samples (instances) in a microarray gene expression dataset. Such structures pose problems to machine learning and make the problem of classification difficult to solve. This is mainly because, out of thousands of genes, most of the genes do not contribute to the classification process. As a result gene subset selection acquires extreme importance towards the construction of efficient classifiers with high predictive accuracy. Random Forests is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages their decisions. The construction is made such that each tree depends on the values of a random vector sampled independently and with the same distribution for all the trees in the forest. The randomness is injected into the tree construction through random split selection, in which the split at each node is decided by a randomly chosen subset of the input features, through random

input selection, in which each tree is grown on a different random subsample of the training data. Machine Learning is considered as a subfield of Artificial Intelligence and it is concerned with the development of techniques and methods which enable the computer to learn. In simple terms development of algorithms which enable the machine to learn and perform tasks and activities. Section II describes the Ensemble classification technique. Section III discusses proposed methodology and Section IV focuses on the experimental results and discussion. Finally, results are summarized and concluded in section V.

II. ENSEMBLING TECHNIQUE

In this section we discuss classification technique of data mining and sampling of imbalance data for classification. Classification is supervised learning process, by the nature of classification are divided into binary classification, rule based classification, multi-class classification, Neural network classification, Machine learning. We present the basic classification techniques. Several major kinds of classification method including decision tree induction, Bayesian networks, rule based classifier, neural network, k-nearest neighbour classifier and fuzzy logic techniques. The ensemble approach to artificial intelligence is a relatively new trend in which several machine learning algorithms are combined [3]. The main idea of the algorithm is to use the strength of a classifier is exciting. Ensembles mainly useful when the problem can be divided into sub problems. In his case, the actors in each module, which may include one or more algorithms assigned to a particular problem.

DECISION TREES: Decision trees are simple knowledge representation and they classify examples to a finite number of classes, the nodes are labelled with attribute names, the edges are labelled with possible values for this attribute and the leaves labelled with different classes. Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes in an object. The following is an example of objects that describe the weather at a given time. The objects contain information on the outlook, humidity etc

NEURAL NETWORKS: Neural network conducts an analysis of the information and provides a probability estimate that the data matches the characteristics which it has been trained to recognize. While the probability of a match determined by a neural network can be 100%, the accuracy of its decisions relies totally on the experience the system gains in analysing examples of the stated problem [6]. A neural network contains no domain knowledge in the beginning, but it can be trained to make decisions by mapping exemplar pairs of input data into exemplar output vectors, and adjusting its weights so that it maps each input exemplar vector into the corresponding output exemplar vector approximately. A knowledge base pertaining to the internal representations is automatically constructed from the data presented to train the network. Well-trained neural networks represent a knowledge base in which knowledge is distributed in the form of weighted interconnections where a learning algorithm is used to modify the knowledge base from a set of given representative cases [12]. A generic form of a neural network intrusion detector is presented in the below Figure. The system use the input labelled data (normal and attack samples) to train a neural network model. The resulting model is then applied to the new samples of the testing data to determine the corresponding class of each one, and so to detect the existing attacks. Using the label information of the testing data, the system can compute the detection performances measures given by the false alarms rate, and the detection rate. A classification rate can also be computed if the system is designed to perform attacks multi classification.

MACHINE LEARNING: Machine Learning is considered as a subfield of Artificial Intelligence and it is concerned with the development of techniques and methods which enable the computer to learn. In simple terms development of algorithms which enable the machine to learn and perform tasks and activities [9]. Machine learning overlaps with statistics in many ways. Over the period of time many techniques and methodologies were developed for machine learning tasks. As regards machines, we might say, very broadly, that a machine learns whenever it changes its structure, program, or data (based on its inputs or in response to external information) in such a manner that it's expected future performance improves. Machine learning usually refers to the changes in systems that perform tasks associated with artificial intelligence (AI).

RANDOM FOREST: The random decision forest concept was first proposed by Tin Kam Ho of Bell Labs in 1995. This model was later extended and formalized by Leo Breiman, who used the more general term Random Forest to describe the overall approach. Breiman demonstrated that RFs are not only highly effective, but they readily address numerous issues that frequently complicate and impact the effectiveness of other classification methodologies leveraged across diverse application domains. In particular, the RF requires no simplifying assumptions. Regarding distributional models of the data and error processes. Moreover, it easily accommodates different types of data and is highly robust to overtraining with respect to forest size.

III. PROPOSED METHODOLOGY

In this paper we proposed a CRBF models are creating for data training for minority and majority class data sample for processing of Random forest classification. The input processing of training phase is data sampling technique for classifier. While single-layer RBF networks can potentially learn virtually any input output relationship, RBF networks with single layers might learn complex relationships more quickly. The function neCrf creates cascade-forward networks. For example, a cascaded layer network has connections from layer 1 to layer 2, layer 2 to layer 3, and layer 1 to layer 3. The cascade-layer network also has connections from the input to all cascaded layers. The additional connections might improve the speed at which the network learns the desired relationship. CRBF artificial intelligence model is similar to feed-forward back-propagation neural network in using the back-propagation algorithm for weights updating, but the main symptom of this network is that each layer of neurons related to all previous layer of neurons. Tan-sigmoid transfer function, log - sigmoid transfer function and pure linear threshold functions were used to reach the optimized status

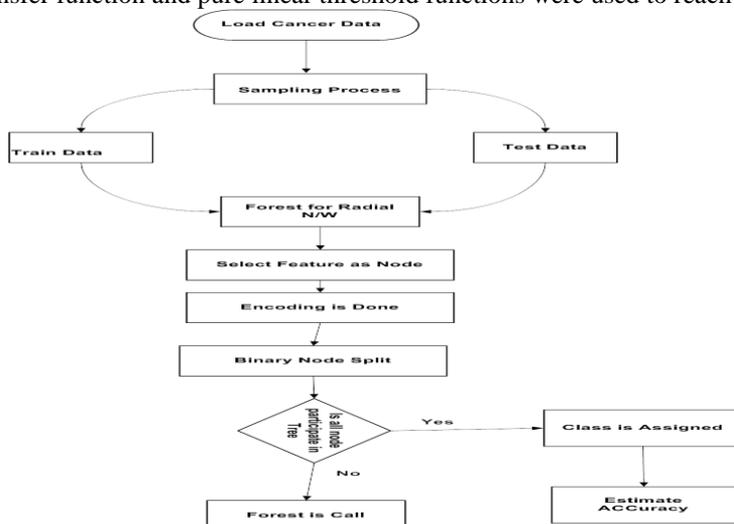


Figure 1: Proposed model for number plate recognition based on feature based optimization

METHODOLOGY STEP

In this section discuss the steps of methodology ensemble and classification. The process of define in following steps.

1. Sampling of data of sampling technique

a) estimate the feature correlation attribute as

$$\text{Rel}(a, b) = \frac{\text{cov}(a,b)}{\sqrt{\text{var}(a) \times \text{var}(b)}} \quad \text{Here } a \text{ and } b \text{ the feature attribute of input data}$$

b) the estimated correlation coefficient data passes through RBF function as

$$x(t) = w_0 + \sum_{j=1}^{\text{total data}} w_j \exp\left(\frac{-(\text{total} - x_j)}{\sigma^2}\right)$$

c) create the relative feature difference value

$$Rc = \sum_{k=1}^r \sum_{i=1}^m (h_i - h)(e_{ik} - e_t)$$

d) After sampling of feature data get reduces set of feature attribute of feature matrix.

2. sampling data passes through downstream and split train data and test data

3. Apply Radial forest for train data for class labelling

4. Apply cross fold ratio constant as 2/3.

For each tree. .

Estimate the classification accuracy and find misclassification of Radial forest

5. For each node in Tree

6. Calculate confusion matrix and estimate Error Rate.

Node selection: if measured Error Rate is decreases class select as major sample.

7. unclassified data passes through Recall forest

IV. EXPERIMENTAL RESULT ANALYSIS

In this section we perform experimental process of proposed classification algorithm for cancer gene classification based on random forest tree ensemble classifier. The proposed method implements in mat lab 7.8.0 and tested with very reputed data set from UCI machine learning research center. In the research work, we have measured classification accuracy, mean absolute error and execution time of classification method. To evaluate these performance parameters I have used six combinational datasets from UCI machine learning repository namely Wisconsin breast cancer (original)-I, WBC-II, WBC-III, WBC-IV, WBC-V and WBC-VI.

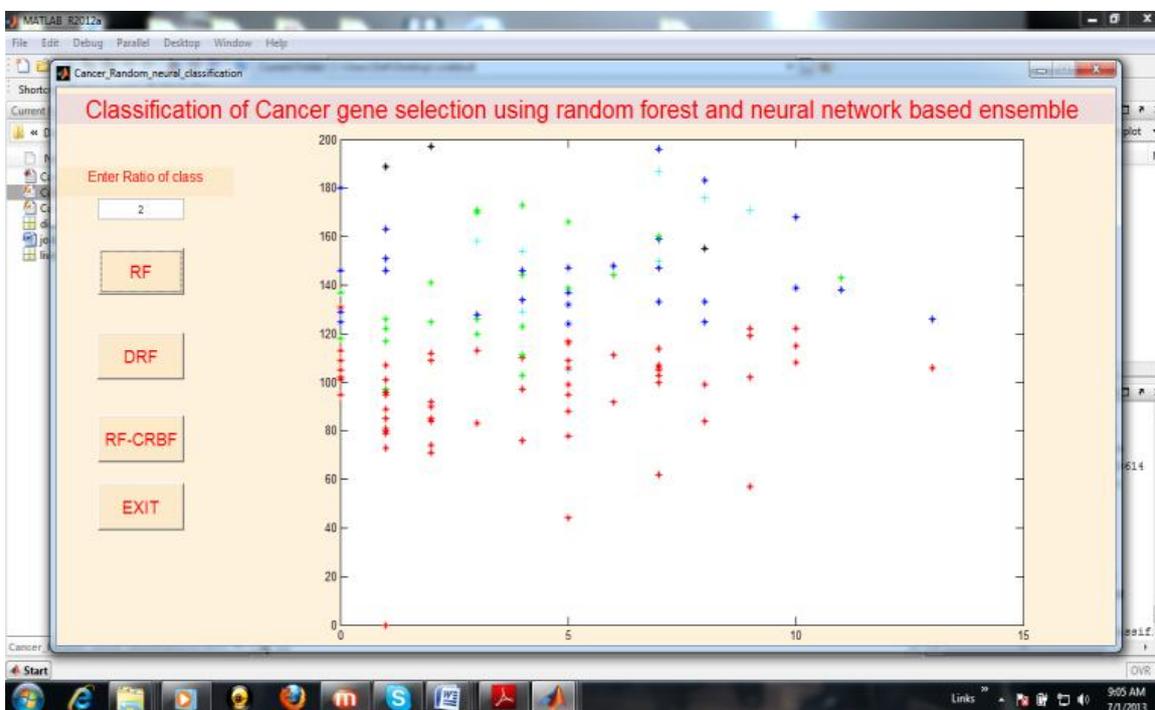


Figure 2: shows that classification of one against all classifier with number of class ratio value 4. The performance of accuracy is 81.54 and means absolute error is 7.35 for cancer data set.

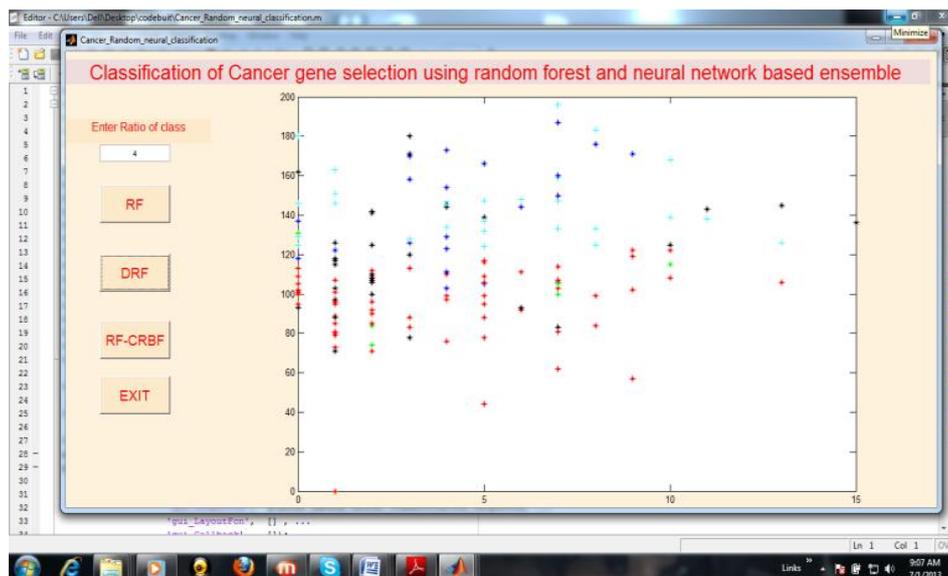


Figure 3: shows that classification of one against all classifier with cascaded RBF number of class ratio value 4. The performance of accuracy is 90.64 and means absolute error is 9.00 for cancer data set.

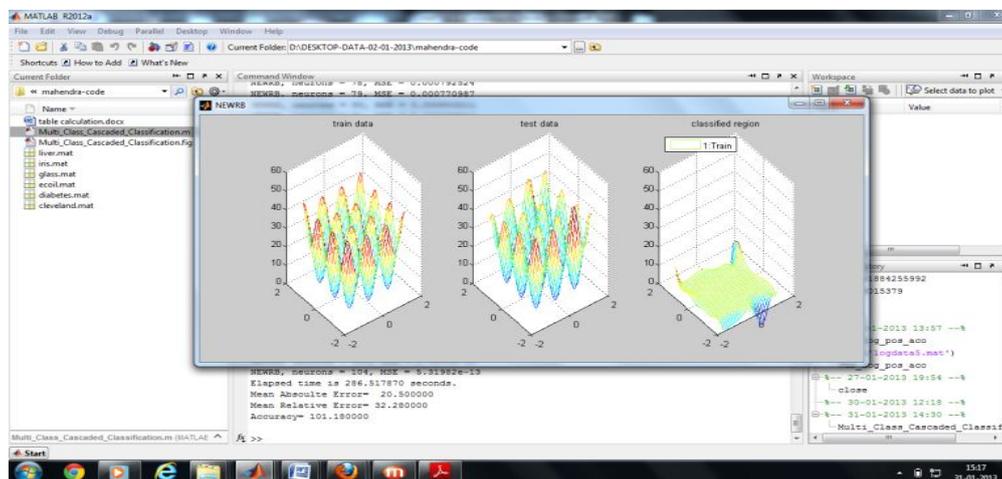


Figure 4: shows that the training of minority gene selection class for data classification.

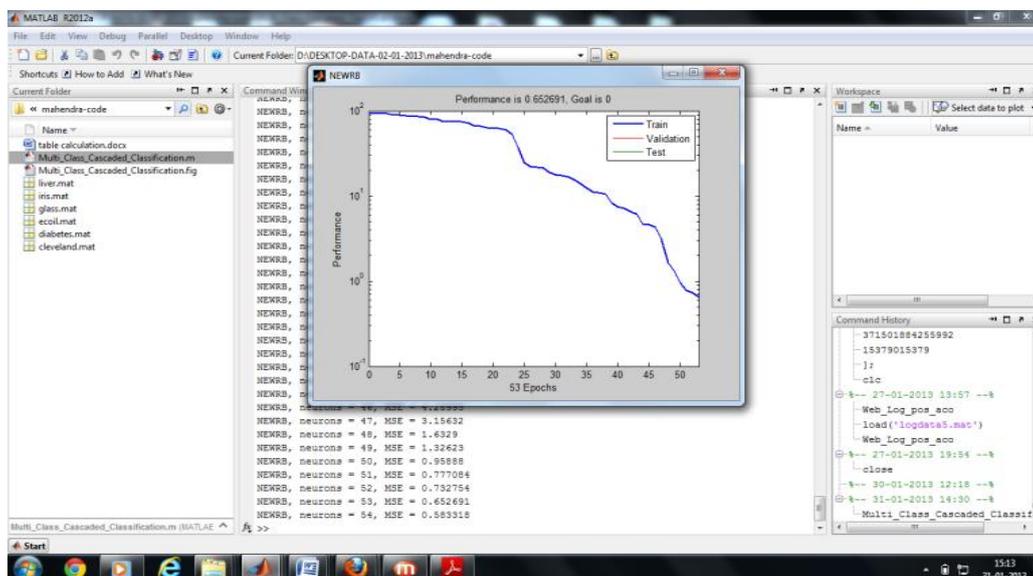


Figure5: shows that the training of minority gene selection class for data classification.

Table 1: Performance evaluation of cancer data set for all classification method.

Ratio of class %	Accuracy			Error		
	RF	DRF	RF-CRBF	RF	DRF	RF-CRBF
20	82.67	87.73	96.86	3.68	5.00	3.50
30	81.70	88.86	97.66	3.12	4.00	2.50
40	82.00	88.68	97.99	3.68	5.00	3.50

Table 2: Performance evaluation of WBC-VI data set for all classification method

Ratio of class %	Accuracy			Error		
	RF	DRF	RF-CRBF	RF	DRF	RF-CRBF
20	82.67	87.73	96.86	3.68	5.00	3.50
30	81.70	88.86	97.66	3.12	4.00	2.50
40	82.00	88.68	97.99	3.68	5.00	3.50

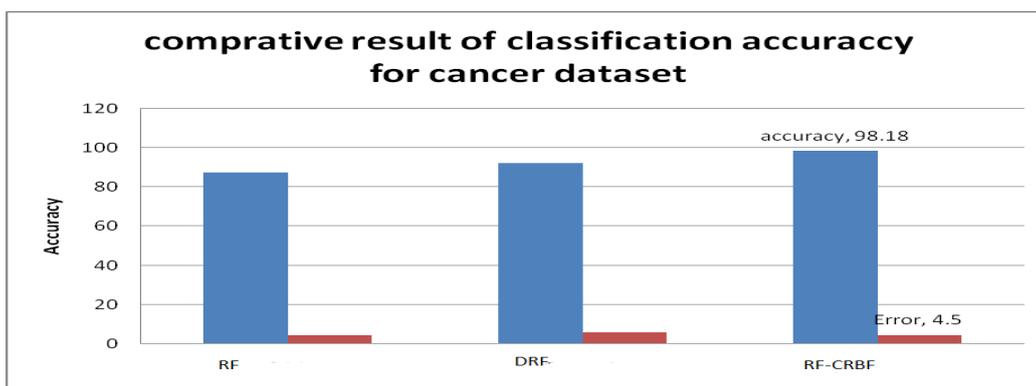


Figure 6: shows that comprative performance analysis of RF, DRF and RF-CRBF method for data balancing for multi-class classification for cancer dataset. Result shows that the data imbalancing factor are reduces, the accuracy of classification are increases.

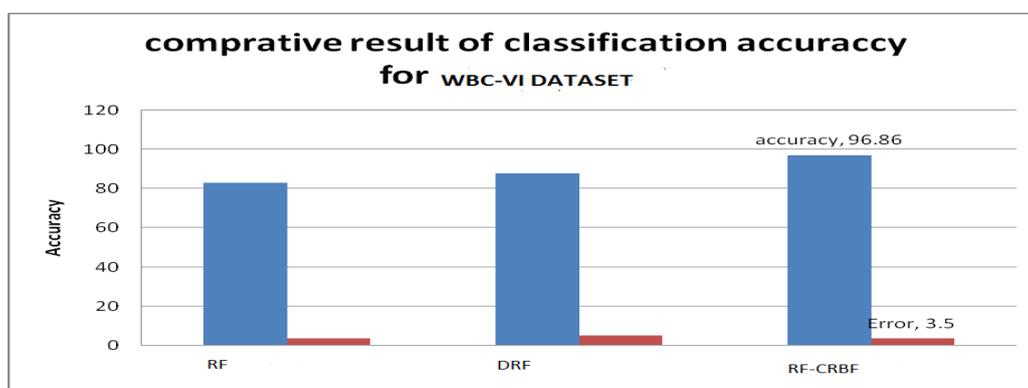


Figure 7: shows that comprative performance analysis of RF, DRF and RF-CRBF method for ensemble classification for WBC-VI dataset. Result shows that the gene selection factor are reduces, the accuracy of classification are increases.

V CONCLUSION AND FUTURE WORK

In this paper we proposed the improved random forest classification technique based on cascaded RBF network. The cascaded RBF network improved the accuracy of minority class of classifier and reduces the unclassified region in random forest classification. The increasing of random forest classification region improved the accuracy and performance of classifier. Our empirical result shows better result in compression of one against all with balanced data in random forest classification. The cascaded RBF network also improved the performance of classifier in terms of complexity of computation.

A technique named as the dynamic random algorithm to solve the random forest imbalanced problem. It applies the binary classification techniques called the DRF approach and the combined data selection technique. Performance of result evaluation shows that our RF-CRBF is better classifier in compression of DRF. In future we used sampling method for the reduction of time and improvement of minority class classification. And another work of future is RBF sampling process applied in SVM random forest classification for better mapping of feature space.

REFERENCES

- [1] Yetian Chen” Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets” in ijcsse-2010
- [2] Suzan Koknar-Tezel and Longin Jan Latecki “Improving SVM Classification on Imbalanced Data Sets in Distance Spaces” Ninth IEEE International Conference on Data Mining,2009
- [3] Haibo He, Member, IEEE, and Edwardo A. Garcia” Learning from Imbalanced Data” in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009.
- [4] Hong Guo and Yi L. Murphey” Neural Learning From Unbalanced Data Using Noise Modeling” in IJCNN , pp.309-314,June 2009.
- [5] Shuiping Gou, Member, IEEE, Hui Yang, Licheng Jiao , Senior Member, IEEE and Xiong Zhuang” Algorithm of Partition based Network Boosting for Imbalanced Data Classification” in IEEE 2010.
- [6] Xue-wen Chen and Michael Wasikowski “FAST: A ROC-based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems” in ijser conference 2010.
- [7] Vaishali Ganganwar” An overview of classification algorithms for imbalanced datasets” in International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4, April 2012.
- [8] S, eyda Ertekin, Jian Huang, Leon Bottou and C. Lee Giles” Learning on the Border: Active Learning in Imbalanced Data Classification” in Journal of Machine Learning Research 6 1579–1619,2005.
- [9] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser, ” svms Modeling for Highly Imbalanced Classification” journal of latex class files, vol. 1, no. 11, november 2002.
- [10] Salvador Garcia, Jose Ramon Cano, Alberto Fernandez and Francisco Herrera “Prototype Selection for Class Imbalance Problems” in Eighth International Symposium on Natural Language Processing, 2009.
- [11] Zhi-Hua Zhou and Xu-Ying Liu “Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem” in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2010.
- [12] Q. Gu, Z. Cai, L. Zhu, and B. Huang, "Data mining on imbalanced data sets," in International Conference on Advanced Computer Theory and Engineering (ICACTE '08), pp. 1020-1024.,2008.
- [13] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, pp. 687-719, 2009.
- [14] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 63-77, 2006.