Classification of Cancer Gene Selection Using Random Forest and Neural Network Based Ensemble Classifier

Jogendra Kushwah[1], Divakar Singh[2]

Abstract

The free radical gene classification of cancer diseases is challenging job in biomedical data engineering. The improving of classification of gene selection of cancer diseases various classifier are used, but the classification of classifier are not validate. So ensemble classifier is used for cancer gene classification using neural network classifier with random forest tree. The random forest tree is ensembling technique of classifier in this technique the number of classifier ensemble of their leaf node of class of classifier. 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. For empirical evaluation of our proposed method we used UCI cancer diseases data set for classification. Our experimental result shows that better result in compression of random forest tree classification.

Keywords

Classification of cancer gene selection, random forests, neural network.

I. Introduction

Microarray gene expression experiments help in the measurement of expression levels of thousands of genes simultaneously. Such data help in diagnosing various types of cancer gene with better accuracy. The fact that this process generates a lot of complex data happens to be its major limitation. 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 ofgenes, 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, and through random input selection, in which each tree is grown on a different random subsample of the training data [4]. One of the most important issues on the creation of ensemble classifiers is the size of the ensemble, the number of the base classifiers which consist of the ensemble. The methods reported in the literature are based on the overproduce-and-choose strategy. The overproduction phase aims of producing a large initial pool of candidate classifiers while the selection phase aims of choosing adequate classifiers from the pool of classifiers so that the selected group of classifiers can achieve optimum positive predictive rate. The methods are differentiated in the second phase (selection phase) where different approaches have been developed. The second solution is a better one and has been considered more than the first approach because binary classifiers are easier to implement and moreover some powerful Algorithms such as Support Vector Machine (SVM) are inherently binary [9]. The complexity of SVM slows down the computation performance. To reduce the complexity due to increase in number of class, the multiclass classifier is simplified into a series of binary classification such as One-Against-One and One-Against-All. In the binary classification there also exists a problem of imbalanced class distribution [12,13]. Using Data Balance algorithm and One against One technique combined this problem is

Jogendra Kushwah, Department of Computer Science & Engineering., BUIT, Bhopal, India.
Divakar Singh, Department of Computer Science & Engineering, BUIT, Bhopal, India.

solved. The imbalanced data problem can be an obstacle for inductive learning machine. So there are two approaches called data level approach and algorithm level approach are there to deal with the data imbalancing problem [15]. The data level approach aims to rebalance the class distribution and is applied before a classifier is trained. While algorithm level approach is used to strengthen the exiting classifier to recognize the small class by adjusting the applied algorithm. The rest of paper is organized as follows. In Section II. Discuss related work of cancer gene classification. The Section III discusses proposed method of cascaded RBF for cancer gene classification IV discusses experimental result.And finally in section V discuss conclusion and future direction of proposed method.

II. Related Work

In this section we discuss cancer gene classification technique related to data imbalance problem.Salvador Garcia, Jose Ramon Cano, Alberto Fernandez and Francisco Herrera entitled a method of Prototype Selection for Class Imbalance Problems as [1] classification algorithms is said to be unbalanced when one of the classes is represented by a very small number of cases compared to the other classes when a set of input is provided. In such cases, standard classifiers tend to be flooded by the large classes and ignore the small ones. Zhi-Hua Zhou and Xu-Ying Liu entitled study of Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem as [2] the effect of sampling and threshold-moving in training cost-sensitive neural networks. These techniques modify the distribution of the training data such that the costs of the examples are conveyed explicitly by the appearances of the examples. PiyasakJeatrakul, KokWaiWong, and Chun Che Fung entitled an analysis of Data Cleaning for Classification Using Misclassification [3] as in most classification or function approximation problems; the establishing of an accurate prediction model has always been a challenging problem. When constructing a prediction model, it is always difficult to have an exact function or separation that describes the relationship between the input vector, X and target vector, Y. Amal S. Ghanem and SvethaVenkatesh, Geoff West entitled problem in Cancer gene Pattern Classification in Imbalanced Data [4] as the majority of cancer gene pattern classification techniques are there for learning from balanced datasets. However, in several real-world domains, the datasets have imbalanced data distribution, where some classes of data may have few training examples compared for other classes. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer entitled a problem of Synthetic Minority Over-sampling Technique [5] as an approach to the construction of classifiers from imbalanced datasets is described. A dataset is imbalanced if the classification categories are not approximately equally represented. Imbalance on the order of 100 to 1 is prevalent in fraud detection and imbalance of up to 100,000 to 1 has been reported in other applications SMOTE provides a new approach to over-sampling. The combination of SMOTE and under-sampling performs better than plain under-sampling. GuobinOu,Yi Lu Murphey entitled an application of Cancer gene pattern classification using neural networks [6] as Cancer gene pattern classification has many applications including text document classification, speech recognition, object recognition, etc. Cancer gene pattern classification using neural networks is not a trivial extension from two-class neural networks Jeatrakul, P. and Wong entitled an Enhancing classification performance of cancer gene imbalanced data using the OAA-DB algorithm [7] as combining the multi-binary classification technique called One-Against-All (OAA)and a data balancing technique. It applies the multi-binary classification techniques called the One-Against-All (OAA) approach and the combined data balancing technique. The combined data balancing technique is the integration of the under-sampling technique using Complementary Neural Network (CMTNN) and the over -sampling technique using Synthetic Minority Over -sampling Technique (SMOTE). Jeatrakul, P., Wong, K.W., Fung, C.C. and Takama entitled a misclassification analysis for the class imbalance problem [8] as the class imbalance issue normally causes the learning algorithm to be dominated by the majority classes and recognize slightly the minority classes. This will indirect affect how human visualize the data. The CMTNN is applied to detect misclassification patterns. For the three techniques I, training data is down sized by eliminating only the misclassification patterns discovered by both the Truth NN and Falsity NN. For technique II, the training data is downsized by eliminating all misclassification patterns discovered by the Truth NN and Falsity NN. These two techniques are applied for under-sampling either the majority class or the minority classes. SofieVerbaeten and Anneleen Van Assche entitled a Methods for Noise Elimination in Classification Problems [9] as ensemble methods combine a set of classifiers to construct a new classifier that is (often) more accurate than any of its component classifiers.

In many applications of machine learning the data to learn from is imperfect. Different kinds of imperfect information exist, and several classifications are given in the literature. He addressed the problem of training sets with mislabeled examples in classification tasks. JareeThongkam ,GuandongXu, Yanchun Zhang and Fuchun Huang entitled a prediction model through improving training space for breast cancer [11] as medical prognoses need to deal with the application of various methods to historical data in order to predict the survivability of particular patients suffering from a disease using traditional analytical applications such as Kaplan–Meier and Cox-Proportional Hazard, over a particular time period. However, more recently, due to the increased use of computing automated tools allowing the storage and retrieval of large volumes of medical data to be collected and made available to the medical research community, there has been increasing interest in the development of prediction models using a new method of survival analysis entitled period analysis. Gustavo E. A. P. A. Batista, Ronaldo C. Prati and Maria Carolina Monard entitled a problem of learning from imbalanced data sets [13] study of the Behavior of Several Methods for Balancing Machine Learning Training Data. There are several aspects that might influence the performance achieved by existing learning systems. It has been reported that one of these aspects is related to class imbalance in which examples in training data belonging to one class heavily outnumber the examples in the other class. In this situation the learning system may have difficulties to learn the concept related to the minority class. P. Jeatrakul and K.W. Wong entitled a problem of Comparing the Performance of Different Neural Networks for Binary Classification [14] as classification problem is a decision making task where manyresearchers have been working on. There are a number of techniques thereto perform classification. Neural network is one of the artificial intelligent techniques that have many successful examples when applying to this problem.In order to solve the classification problems and prediction, many classification techniques has been proposed. Some of the successful techniques are Artificial Neural Networks (ANN), Random forests (RF) and classification trees. Jose G. Moreno-Torres and Francisco Herrera entitled a Study on Overlapping and Data Fracture in Imbalanced Domains by means of Genetic Programming-based Feature Extraction [15] author proposed GP-RST, a GP-based feature extractor that employs RST techniques to estimate the fitness of individuals. We found GP-RST to be a competitive preprocessing method for highly imbalanced datasets, with the added advantage of providing bio dimensional representations of the datasets it preprocesses, which are easily interpreted.

III. Proposed Method

Artificial neural networks (ANNs) proved to be capable of finding internal representations of interdependencies within raw data not explicitly given or even known by human system. Its special characteristic together with the simplicity of building and training ANNs and their very short response time encouraged their application to the task of feature selection. Because of their inherent nonlinearity, ANNs are able to deal with the complex interactions between variables that affect feature selection. There is no need for complex functional models to describe the relationships between the input variables and the input image. Cascading of neural network model play an important role in data classification and feature optimization. In this section we discuss cascaded model of RBF network for cancer gene classification. The great advantage of RBF network is single layer processing unit and target output independent with input data. In the process of cascading input feature passes through margin of classifier, margin classifier function separate data into layers such as positive and negative in data space domain. The part of positive and negative used as input in cascaded model. We applied to train a neural network to learn the classification features from the data samples of a minority class in the training set and to make more favorable decisions to the minority class. For the convenience of description, we referred to the two classes of data as minority and majority classes respectively. In many applications, if error is inevitable, a neural network is expected to err on one particular class rather than the other The algorithm we investigated was to generate new minority data samples near the classification boundary using the Gaussian and add these new data samples to the training data. The neural networks trained on this set should make more favorable decision to the minority class with the minimization of misclassification of the majority class and have increased generalization capability. For every data sample s of the minority class in the training set, we attempt to generate p new data samples around s subject to its localGaussian distribution of the opposite class. Let us assume the input vector is M dimensional.

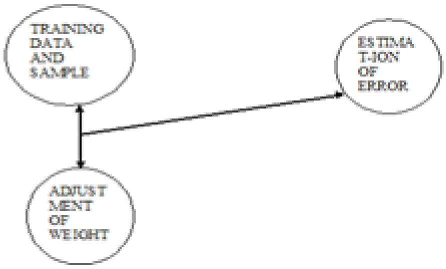

Figure 1: shows the training and adjustment of RF during RBF and estimate error rate.

IV. Experimental Result

In this section we discuss the modification of our cascaded model over random forest classifier. The basic classifier used as decision tree. The split of tree is replaced with cascaded RBF kernel function. It also called two kernel function of classification. For the performance evaluation we used dataset forms UCI machine learning respositry. These datasets are cancer, dataset are used. Our modified classifier implements in matlab 7.8.0 software package and used library function of Random forest. Here we show some classified data region using RF and cascaded RBF model.

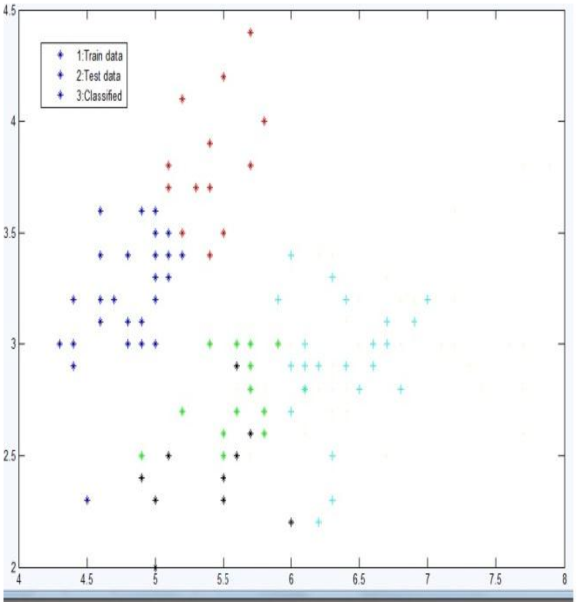

Figure 2: shows that classification process of Random forest with cascaded RBF network. In this figure shows that three region of data train data, test data and finally classified data.

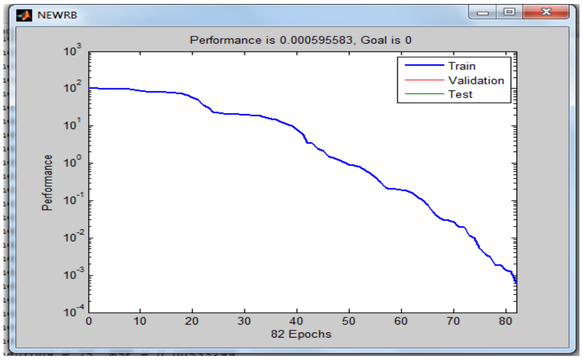

Figure 3: shows that the train model process of RBF network, in this model network we used 300 neurons data point and 100 epochs cycle for train of data.

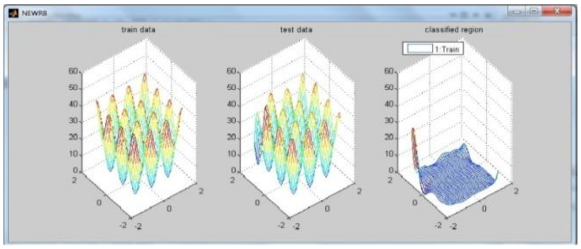

Figure 4: shows that the data classification region in cascade mode with Random forest, all data passes through model and acquired sample color mapped according to data region and finally shows that unclassified region of cancer gene classification.

The empirical evaluation of data in all dataset shows against with method and find accuracy and error rate basis on class ratio of classifier. All these method such as RF, DRF and RF-CRBF as tabular form. For Cancer Data, all data find accuracy and error show that performance of classifier all these shows that performance valuation of random forest, random forest neural network data and finally RF cascade RBF network.

Table 1: Comparison

Ratio of class (in %)	Accuracy			Error		
	RF	DRF	RF	DRF	RF-	CRBF
30	87.08	92.18	98.18	4.56	6.00	4.50
40	87.08	92.18	98.68	3.92	5.00	3.50
50	87.08	92.18	98.07	3.53	5.00	3.50

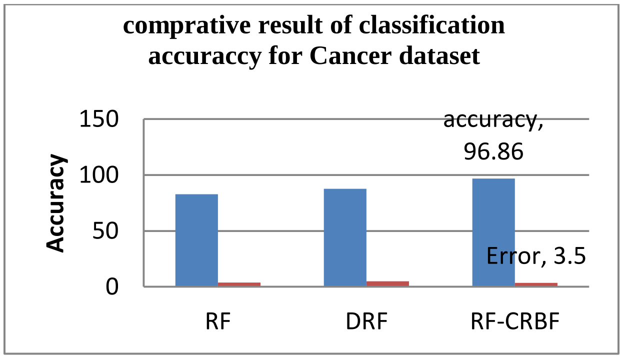

Figure 5: shows that the data classification analysis of random forest and D random forest and random forest with neural network the classification accuracy shows that better accuracy in random forest with neural network.

V. Conclusion and Future Work

In this paper we discuss the improved cancer gene classification technique based on cascaded RBF network. The cascaded RBF network improved the accuracy of minorityclass of classifier and reduces the unclassified region in cancer gene classification. The increasing of cancer gene classification region improved the accuracy and performance of classifier. Our empirical result shows better result in compression of RF with balanced data in cancer gene classification. The cascaded RBF network also improved the performance of classifier in terms of complexity of computation. We showed through experimental results that the cascaded algorithm is effective in the training of both RF and neural networks. We speculate that the algorithm can be extrapolated to the general classification problem of P classes within which the p classes are to be emphasized, where p <P. By generating noise data samples along the classification boundaries for these p classes using the noise cascaded algorithm, the trained neural network would have increased classification capability and generalization ability over the p classes.

References

[1] P. Jeatrakul and K.W. Wong “Comparing the Performance of Different Neural Networks for Binary Classification Problems” in Eighth International Symposium on Natural Language Processing, 2009.

[2] 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, 2006.

[3] Piyasak Jeatrakul, KokWaiWong, and Chun Che Fung “Data Cleaning for Classification Using Misclassification Analysis” in Data Cleaning for Classification Using Misclassification Analysis, 2010.

[4] Amal S. Ghanem and SvethaVenkatesh, Geoff West “Cancer gene Pattern Classification in Imbalanced Data” in International Conference on Pattern Recognition, 2010.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer “SMOTE: Synthetic Minority Over-sampling Technique” in Journal of Artificial Intelligence Research 16 (2002) 321–357, 2002.

[6] GuobinOu,Yi Lu Murphey “Cancer gene pattern classification using neural networks” in The Journal of the Pattern Recognition Society, 2007.

[7] Jeatrakul, P., Wong and K.W. “Enhancing classification performance of cancer gene imbalanced data using the OAA-DB algorithm” in Annual International Joint Conference on Neural Networks (IJCNN), 2012.

[8] Jeatrakul, P., Wong, K.W., Fung, C.C. and Takamain“Misclassification analysis for the class imbalance problem” INWorld Automation Congress (WAC), 2010.

[9] SofieVerbaeten and Anneleen Van Assche” Ensemble Methods for Noise Elimination in Classification Problems”IEEE Transaction, 2003.

[10] Jeatrakul, P., Wong, K.W. and Fung “Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm” (ICONIP), 2010.

[11] Jaree Thongkam ,Guandong Xu, Yanchun Zhang, Fuchun Huang “Toward breast cancer survivability prediction models through improving training space” in Expert Systems with Applications, 2009.

[12] Jeatrakul, P., Wong, K.W. and Fung “Using misclassification analysis for data cleaning” in International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII), 2009.

[13] Gustavo E. A. P. A. Batista, Ronaldo C. Prati and Maria Carolina Monard “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data” inSigkdd Explorations, 2004.

[14] P. Jeatrakul and K.W. Wong “Comparing the Performance of Different Neural Networks for Binary Classification Problems” in Eighth International Symposium on Natural Language Processing, 2009.

[15] Jose G. Moreno-Torres and Francisco Herrera “A Preliminary Study on Overlapping and Data Fracture in Imbalanced Domains by means of Genetic Programming-based Feature Extraction” in IEEE Transaction,2010.