IMPROVING THE CLASSIFICATION PERFORMANCE OF MULTI-CLASS IMBALANCE DATA USING THE RADIAL BIASED NEURAL NETWORK FUNCTION

Shraddha Patil[1], Prof.Piyush Singh[2]

[1]PG Scholar, [2]Guide
Department of Computer Science & Engineering,
R.K.D.F. Institute of Science & Technology, Bhopal, India

ABSTRACT

Multi-class classification is extended form of binary classification. Binary classifier such as support vector machine ,nearest neighbor and decision tree impart a role of multi-class classification. In multi-class classification task of classification are performed by two different method such as one against one(OAO) and one against all(OAA). In process of such method featuer selection play important role in classification. The proper selection of feature depends on data blancing. Various authors and research modified the multiclass classification approach such as one against one and one against all. In both method OAO and OAA create a unclassified region for data and decrease the performance of classifier such as support vector machine. For the reduction of unclassified region we used feature reduction technique using radial basis function network. In this paper we modified the sampling technique for imbalance data classification using multi-class 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 macine learning respositry. Our proposed method implement in matlab 7.8.0

Keywords -- Data Mining, SVM, RBF

I. INTRODUCTION

Unbalanced dataset learning is a new paradigm of machine learning which has applicability in real time, since all the datasets of real time are of unbalanced nature. For example if you consider a case of medical surgery or scientific experimental analysis the cases in this study will not be of balanced category. The cases may be more in either positive category or negative category, thereby creating an unbalanced dataset. Unbalanced data set is that samples of some classes in the data set are more than samples of other classes, classes with more samples are called majority class, and on the other hand, classes with a few samples are called minority class [1]. In the case of unbalanced datasets the common shortcoming using traditional classifiers is that they misclassify minority dataset as majority dataset. In real time scenario this misclassification will cost a lot to the area of applicability in terms of money if it is banking domain, in terms of life if it is medical domain, in terms of quality if it is quality control domain etc. In real time domain the classification accuracy of minority class is as

equal as classification accuracy of majority class. There is an urgent need to improve the classificatory performance of minority class in the fields of machine learning and pattern recognition [2]. Imbalanced data classification often arises in many practical applications. Many classification approaches are developed by assuming the underlying training set is evenly distributed. The data imbalance is another problem existed in the multiclass classification for machine learning. In the process of machine learning (ML) if the ration of minority class and majority class is highly different then machine learns more by the majority class and learns less from minority class. So to resolve the issue of minority class in imbalance data set special characteristic technique is required[3]. Two more common approaches to solve this problem are data level approach and algorithm level approach. Data level approach rebalances the data before a classifier is trained and algorithm level approach strengthens the classifier. Alternatively, the learning algorithm can be modified to account for class imbalance. In this section we discuss cascaded model of RBF network for multi-class 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 [4]. 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 error 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 favourable decision to the minority class with the minimization of misclassification of the majority class and have increased generalization capability [7]. 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 local Gaussian distribution of the opposite class. Section-I gives the introduction of the data balancing. Section-II cascaded RBF function. multi-class classification technique in III. In section IV proposed algorithm. in section V discuss comparative result with standard parameter. Finally, in section-VI conclusion and future scope.

II CASCADED RADIAL BAISE FUNCTION NETWORK (RBF)

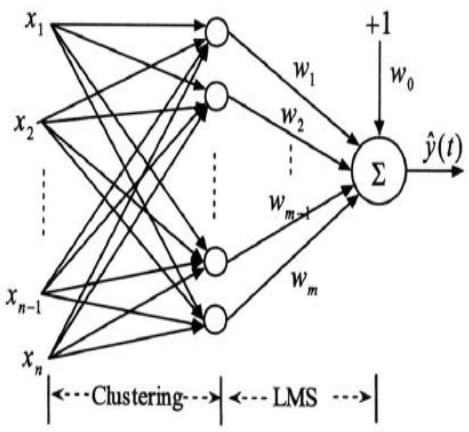

Figure 1: shows that neuron structure of CRBF

A CRBF Network which is linear in the parameters provided all the RBF centers are prefixed. Given fixed centers i.e. no adjustable parameters the first layer or the hidden layer performs a fixed nonlinear transformation, which maps the input space onto a new space [15]. With n inputs and m hidden neurons is shown in Fig. 1.

Such as network can be representd as

$$\hat{y}(t) = w_0 + \sum_{j=1}^{m} w_j \exp\left(-\frac{\|x - x_{cj}\|^2}{\sigma_j^2}\right) \quad \ldots\ldots\ldots\ldots 1$$

In which x is the input vector, xc; is the center and a2 is the width of the j th Gaussian hidden unit. Parameters wj are weights of the connections that feed the output unit. The output layer then implements a linear combiner on this new space and the only adjustable parameters are the weights of this linear combiner. These parameters can therefore be determined using the linear least mean square algorithm (LMS), which is an important advantage of this approach. This is basically how an RBF network works. Our CCRBF networks are as follows. The training begins with minimal structure (no hidden units), and then more connections, neurons are added to the network according to some redeemed rule. Adding the hidden units one by one is started by first creating a set of candidate hidden units, which can be done in many existent ways. Also, at this stage all the inputs are connected to all the candidate hidden units. Then cascade correlation criterion value is calculated for each candidate hidden unit as follows:

$$C = \sum_{k=1}^{r} \left| \sum_{i=1}^{m} (h_i - \bar{H})(e_{i,k} - \bar{E}_k) \right| \quad \ldots\ldots..2$$

Where hi is the candidate unit output for i th training pattern, eik is the output error of the k th output unit, H is the mean of hi's and Ek is the mean of eik. The candidate hidden unit which gives the largest value for this criterion is installed to the network. The selected unit has the important property that its output values correlate most strongly (compared with other candidate units) with the residual error of the network outputs. The installation of the selected unit is done simply by connecting its output to all the network output units. Hereafter, the parameters of the new unit are frozen. Following the installation of the new unit, all the connections feeding the output units are trained to minimize the error of the network outputs.

III OAA-DB

The One-Against-All technique with Data Balancing (OAA-DB) algorithm is deal with the multi-class classification with imbalanced data. The fundamental principles under this approach are based on the attempt to balance data among classes before performing multi-class classification. The OAA-DB approach combines the OAA and the data balancing technique using the combination of SMOTE and CMTNN. The OAA-DB

technique is an extended algorithm from the OAA. It aims to improve the weakness of OAA because OAA has highly imbalanced data between classes when one class is compared with all the remaining classes [12]. Moreover, if OAA uses only the highest output value to predict an outcome, there is a high potential risk that the majority class can dominate the features of the prediction. The concept of codeword which is used also applied to this OAA-DB technique in order to define the confidence value of the prediction outcomes. The data balancing technique which combines of CMTNN and SMOTE, and followed by the algorithm of OAA-DB the purpose of the OAA-DB algorithm aims to reduce the ambiguity problem of the OAA approach. This is because the OAA approach consists of *K* binary classifiers and they are trained separately. This can cause the classification boundary to be drawn independently by each classifier as shown in Figure 2. The OAA approach may not generalise well on the test data. In this case, the confident bit of codeword and the data balancing technique with the OAA-DB algorithm are used in order to reduce these problems. The confident bit of codeword can be used to decide a class label with confidence at the overlapped region. The data balancing technique also aims to reduce the problem at the uncovered region.

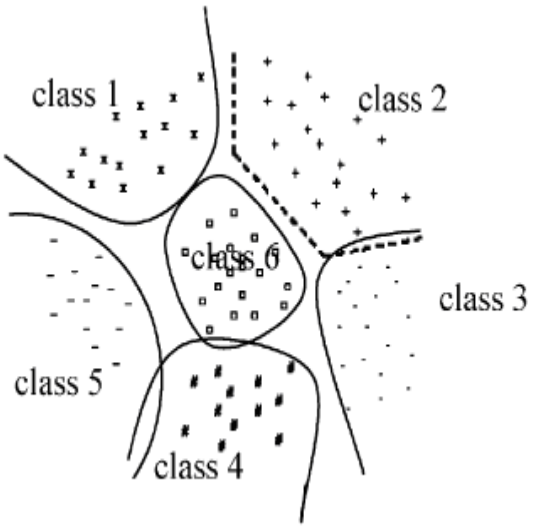

Figure 2: Classification boundaries drawn by classifiers trained with the OAA approach

IV PROPOSED METHODOLOGY

In this section we discuss cascaded model of RBF network for multi-class 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 error 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 local Gaussian distribution of the opposite class. Let us assume the input vector is M dimensional

1. Data are divvied into training and test
2. The training phase data are passes through SMOTE AND CMTNN sampler
3. The sampling of data passes through CRBF AND balanced the data for minority and majority ratio of class
4. The sampled data assigned to k-type binary class
5. Binary class data are coded in bit form
6. if code bit value is single assigned the class value
7. Else data goes to training phase
8. Balanced part of training is updated

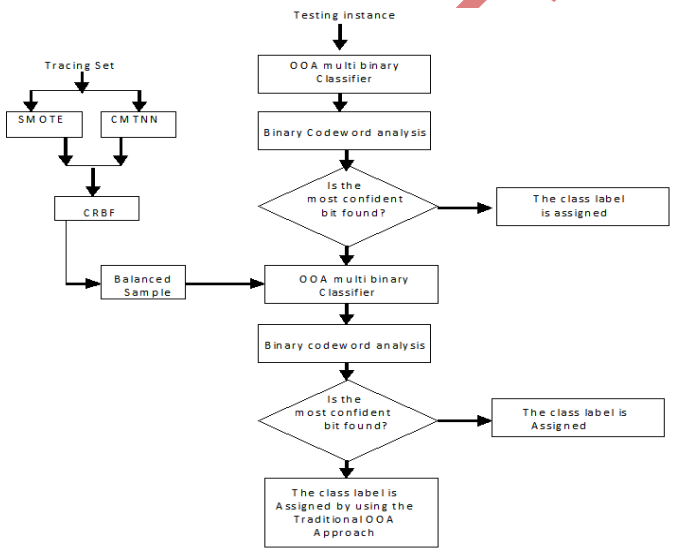

Figure 3: process block diagram of modified OAA-CRBF

V EXPERIMENTAL RESULT ANALYSIS

In this section we discuss the modification of our cascaded model over binary classifier. The basic classifier used as support vector machine. The kernel of support vector machine is replaced with cascaded RBF kernel function. It also called two kernel function of classification. For the performance evaluation we used six dataset forms UCI machine learning respositry. These datasets are cancer, glass, iris, page ,yeast and finally wine dataset are used. Our modified classifier implements in matlab 7.8.0 software package and used library function of support vector machine. Here we shows some classified data region using SVM and cascaded RBF model.

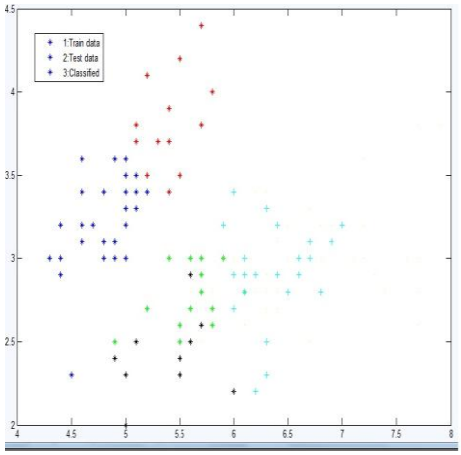

Figure 4 shows that classification process of support vector machine with cascaded RBF network. In this figure shows that three region of data train data, test data and finally classified data.

The empericla 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 OAA, OAA-DB and OAA-CRBF as tabuler form.For Cancer Data, glass data and all data find accuracy and error show that perforance of classifier.

Table 1 gives the classification accuraacy and error rate of cancer data

Ratio of class (in %)	Accuracy			Error		
	OAA	OAA-DB	OAA-CRBF	OAA	OAA-DB	OAA-CRBF
30%	87.08	92.18	98.18	4.56	6.00	4.50
40%	87.08	92.18	98.56	3.92	5.00	3.50
50%	87.08	92.18	99.32	3.53	5.00	3.50

Table 2 gives the classification accuraacy and error rate of iris data

Ratio of class (in %)	Accuracy			Error		
	OAA	OAA-DB	OAA-CRBF	OAA	OAA-DB	OAA-CRBF
20%	80.90	86.00	95.00	2.75	4.00	2.50
30%	80.90	86.00	95.00	2.50	4.00	2.50
40%	80.90	86.00	95.00	2.37	3.00	1.50

Table 3 gives the classification accuraacy and error rate of glass data

Ratio of class (in %)	Accuracy			Error		
	OAA	OAA-DB	OAA-CRBF	OAA	OAA-DB	OAA-CRBF
30%	82.22	87.32	96.32	2.94	4.00	2.50
40%	82.22	87.32	96.32	2.70	4.00	2.50
50%	82.22	87.32	96.32	2.56	4.00	2.50

Table 4 gives the classification accuraacy and error rate of yeast data

Ratio of class (in %)	Accuracy			Error		
	OAA	OAA-DB	OAA-CRBF	OAA	OAA-DB	OAA-CRBF
20%	82.85	87.95	96.95	3.72	5.00	3.50
30%	82.85	87.95	96.95	3.15	4.00	2.50
40%	82.85	87.95	96.95	2.86	4.00	2.50

All these shows that performance valuation of one against all, one against all with balance data and finally one against all cascade RBF network

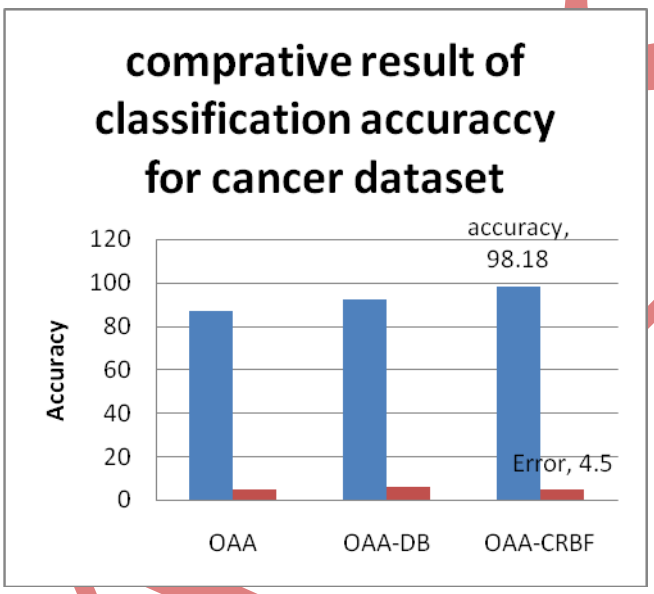

Figure 5: shows that comprative performance analysis of OAA, OAA-DB and OAA-CRBF method for data balancing for multi-class classification for cancer dataset. Result shows that the data imbalancing factor are reduces, the accuracy of classification are increases.

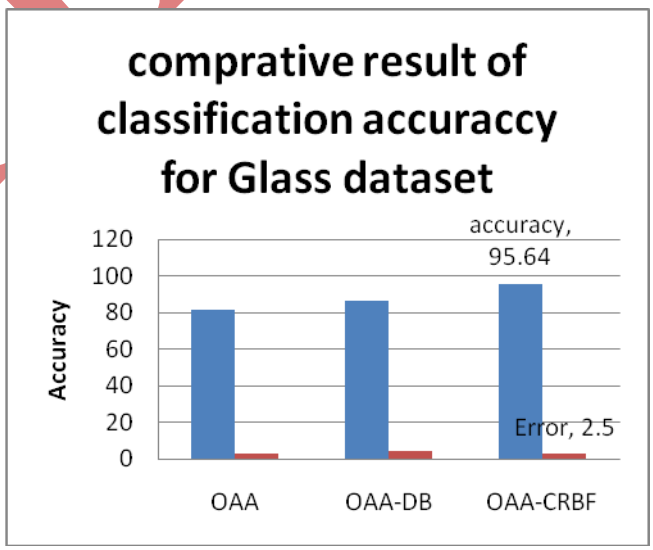

Figure 6 shows that comprative performance analysis of OAA, OAA-DB and OAA-CRBF method for data balancing for multi-class classification for Glass dataset. Result shows that the data imbalancing factor are reduces, the accuracy of classification are increases.

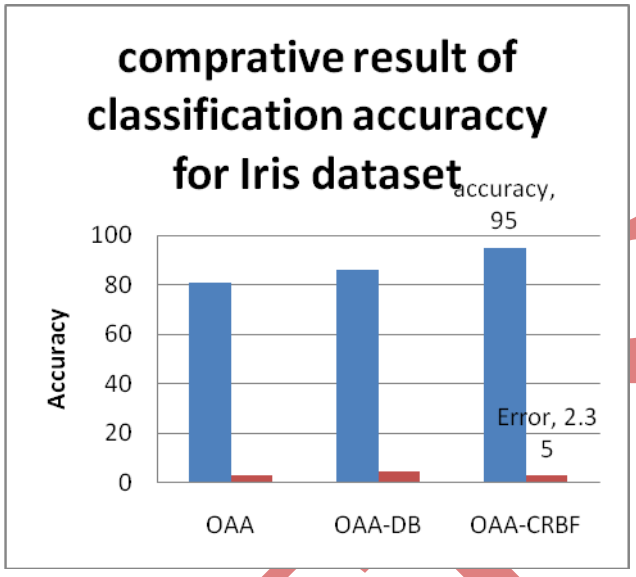

Figure 7: shows that comprative performance analysis of OAA, OAA-DB and OAA-CRBF method for data balancing for multi-class classification for Iris dataset. Result shows that the data imbalancing factor are reduces, the accuracy of classification are increases.

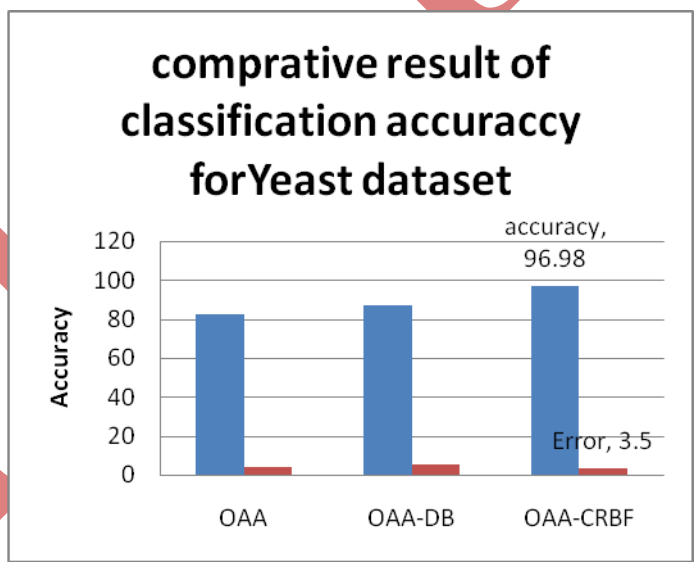

Figure 8: shows that comprative performance analysis of OAA, OAA-DB and OAA-CRBF method for data balancing for multi-class classification for Yeast dataset. Result shows that the data imbalancing factor are reduces, the accuracy of classification are increases.

VI CONCLUSION AND FUTURE WORK

In this paper we discuss the improved multi-class 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 multi-class classification. The increasing of multi-class 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 multi-class 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 SVM 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.

[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 Svetha Venkatesh, Geoff West "Multi-Class 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, 321–357, 2002.

[6] Guobin Ou,Yi Lu Murphey "Multi-class 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 multi-class 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 Takama in"Misclassification analysis for the class imbalance problem" INWorld Automation Congress (WAC), 2010.

[9] Sofie Verbaeten and Anneleen Van Assche" Ensemble Methods for Noise Elimination in Classification Problems" in IEEE Transaction.

[10] Jeatrakul, P., Wong, K.W. and Fung "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm" in 17th International Conference on Neural Information Processing (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" in Sigkdd Explorations.

[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.