A Novel Multi-Class Ensemble Model for Classifying Imbalanced Biomedical Datasets

ThulasiBikku[1],DrN Sambasiva Rao[2], DrAnanda Rao Akepogu[3]

CSE Department, Vignan'sNirula Institute of Technology and Science for Women, Palakaluru, A.P, India[1]
Principal, SRITW, Warangal, Telangana, India[2]
Director of Academics & Planning, JNTUCEA, Ananthapuramu, India[3]

Abstract: This paper mainly focuseson developing aHadoop based framework for feature selection and classification models to classify high dimensionality data in heterogeneous biomedical databases.Wide research has been performing in the fields of Machine learning, Big data and Data mining for identifying patterns. The main challenge is extracting useful features generated from diverse biological systems. The proposed model can be used for predicting diseases in various applications and identifying the features relevant to particular diseases. There is an exponential growth of biomedical repositories such as PubMed and Medline, an accurate predictive model is essential for knowledge discovery in Hadoop environment. Extracting key features from unstructured documents often lead to uncertain results due to outliers and missing values. In this paper, we proposed a two phase map-reduce framework with text preprocessor and classification model. In the first phase, mapper based preprocessing method was designed to eliminate irrelevant features, missing values and outliers from the biomedical data. In the second phase, a Map-Reduce based multi-class ensemble decision tree model was designed and implemented in the preprocessed mapper data to improve the true positive rate and computational time. The experimental results on the complex biomedical datasets show that the performance of our proposed Hadoop based multi-class ensemble model significantly outperforms state-of-the-art baselines.

Keywords: Ensemble model, Map-Reduce, Medical databases, Bioinformatics, Textual Decision Patterns.

I. INTRODUCTION

In the area of data mining and machine learning the feature selection plays a vital role in reducing the complexity while improving the accuracy and computational speed of a classification algorithm [1, 3].The performance of the classifier mainly degraded not only by

the irrelevant features and high dimensional data, but also increases the misclassification rate and storage space especially in imbalance databases.

In order to perform classification, a predictive model is learned from training data. A model is an internal representation of the prediction problem that allows us to classify previously unseen samples. Data Classification became very difficult due to heterogeneous databases and large volumes of data. Data mining and machine learning researchers have identified a major problem,i.e.class imbalance problem, itoccurs when one class instance,are overwhelmed (majority) by another class instance (minority). The class imbalance problem arises when the class of interest is relatively rare and has fewerinstances compared to the majority class [2].

Most of the algorithms are concentrating on classification of data in majority class, whereas ignoring or misclassifying the minority class. Class Imbalance property occurs in medical documents when one or more of the genes or diseases contain more samples known as a majority class in comparison to minority class. In this type of situations, most of the classifier isinfluenced towards the majority classes and hence show very meagerperformance rates on minority classes. It is also possible that classifier predicts everything as major class and ignores the minor class. Many algorithms have been proposed in order to solve the problems connected with class imbalance.Generally, the algorithms are divided into three basic categories:data-preprocessing, the algorithmic approach, and feature selection approach.

In data preprocessing technique, sampling of data is applied, where new samples are added or existing samples are removedto analyze the data. The process of adding new sample in existing minority class is known as oversampling and removing a sample from majority class is known as undersampling. The ensemble is an efficient model that combines different classification methods in Hadoop framework so as to optimize the overall accuracy and true positive rate. Using a specific traditional ensemble model is difficult to handle all kinds of genes or diseases in Hadoop framework. Moreover, the true positive rate of the different classifiers may vary in due to different feature datasets [2]. Selection ofthe most relevant features and high dimensional features improves the classifier’s accuracy and performance.For example,in the case of breast cancer cell recognitionin medical diagnosis, misclassifying non-cancerous cells may lead to some supplementary clinical testing but misclassifying breast cancerous cells lead to serious health risks. However, due to the classification of imbalanced datawhich leads to the class imbalance problem, where the minority class data are more likely to be misclassified than the

majority class data, due to the design principles used in the algorithms.Most of the data mining and machine learning algorithms optimize the overall performance of classifier and classification accuracy which results in the misclassification of minority class data.

In our proposed research work, we use a heterogeneous multi-class ensemble decision tree model based on naïve Bayesian tree classification. In biomedical databases, the research is increasing on the genes or diseases and the database of genes is growing exponentially, which requires parallel computations to handle the high dimensional features [4, 5]. This motivates us to implement a model for both feature selection and multi-class ensemble model in the Hadoop framework.

The rest of the paper contains the related work of the different classification models in the big data framework are discussed in Section II. In section III, explains about the proposed model a novel ensemble model for classifying and pattern analysis using the Hadoop framework. In Section IV, experimental results are evaluated on different medical data sets taken from different repositories like MEDLINE and PubMed and finally, Section V discusses about conclusion of the proposed algorithm and future scope of the proposed model.

2. RELATED WORK

Feature selection is a vitalproblem in machine learning and data mining. It aims to select relevant features that improve the accuracy and performance of the classifier. High dimensional data and irrelevant features may reduce the performance of the classifier and increase the storage space misclassification rate especially in imbalance databases [5]. The categories of Feature selection metrics are: one-sided or two-sided based on positive features they select or combination of positive and negative features of the data considered [6, 7]. Depending on the data type, feature selection metrics can be categorized as binary or continuous. For example: Chi square(CHI), Mutual Information (MI), Information Gain (IG) and Odds Ratio (OR) can handle both ordinal and nominal data (binary). But Pearson correlation coefficient, Feature Assessment by Information Retrieval(FAIR), Feature Assessment by Sliding Threshold (FAST) and Signal to Noise ratio (S2N) can handle continuous data [9]. Binary metrics performance completely depends on the change in threshold value.Nguwi and Cho [8] discussed a feature selection method derived from SVM, used ranking criteria and eliminated less contributing features. Alibeigi et al. [9] presented a feature ranking approachusing probability density estimation for small size ofthe sample and high dimensional data sets. The

class imbalance problem can be done by adjusting the prior distribution for minority and majority class and constructed balanced training data set. Mazurowskia et al. [12] concluded that particle swarm optimization (PSO) was more responsive to the class imbalance problem, the size of the training sample and high dimensionality features. Li et al. [13] used granular support vector machines to improve the efficiency of the classifier and reduced the computational cost of the algorithm. Yen and Lee [14] proposed asan approach where training data were divided into a small number of clusters and representative data samples for majority class were selected from each cluster based on the ratio of majority class samples to minority class samples. Kang and Cho [15] proposed an ensemble of under sampled support vector machines (EUS SVMs) to overcome the information loss due to under sampling method and reduced the time complexity of oversampling method due to artificial data. Guo and Viktor [10] proposed a DataBoost-IM, which is an ensemble based learning approach, which is used for data generationand classification of imbalanced data. Zhang and Luo [11] proposed a parallel classification method to speed up the classification of features, but in some cases the acceleration is not enhanced. Lia and Suna [16] proposed an ensemble with nearest neighbor (NN) and a bagging and solved low classification problems.

Schapire et al[17] presented the first practical boosting algorithm named as AdaBoost algorithm, to overcome the class imbalance problem improved prediction of minority class, but it ignores overall performance of the classifier. Chris Seiffert[18] presented a RUSBoost algorithm, which is a combination of sampling and boosting.This algorithm is used for obtaining patterns from skewed training data and alternative to SMOTEBoost, which is a combination of boosting and sampling algorithms. The main drawback of this algorithm is, it is unable to solve multiclass imbalance problem. Art B. Owen[19] considers the data which is infinitely imbalanced i.e. one class has a finite size of the sample and the other class size of sample grows without limit. Here he presented logistic regression provided a solution to the infinitely imbalanced case. Mostly it is used for binary classification, but the algorithm performance depends on the number of outlier in the data. Glenn Fung and Olvi L. Mangasarian [20] presents Proximal Support Vector Machines (PSVM),which assigns points to the closest of two parallel planes to classify, which are used to handle the class imbalance problem, whereas the distribution of the sample is not considered. Elkin García and Fernando Lozano [21] discussed a classification algorithm based on Support Vector Machines combined with

Boosting techniques improved the performance of the classifier to predict minority sample and class imbalance distribution is overlooked.

Many areas like machine learning and data mining are affected by class imbalance problems, sothe solution provided by many algorithms in data mining is helpful but not enough due to the large volumes of data. The technique is considered for handling a problem of distribution of data is extremely depends upon the type of data used for research.

3. PROPOSED MODEL

In this proposed system, biomedical documents are collected from the PubMed/MEDLINErepositories in the XML format. In this model, a novel biomedical MapReduce framework was used to discover the textual relationships from a large collection of medical document sets.

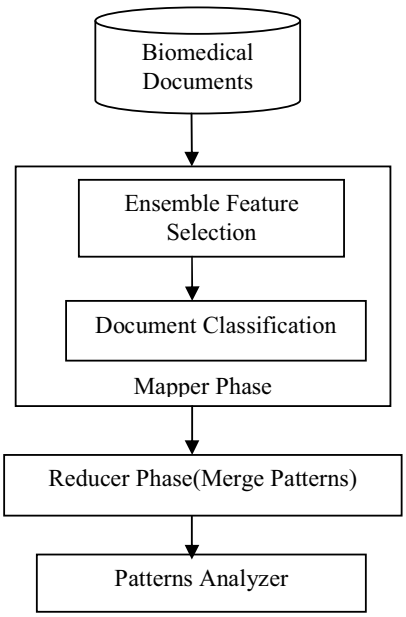

Figure 1: Proposed Multi-class ensemble classifier using hash join operation

Each document in the document set is filtered using the ensemble feature selection measures.Each filtered feature selection documents is classified using enhanced multi-nominal naïve Bayesian model as shown in Figure 1.

Mapper Phrase

Input: Biomedical Documents D, Threshold λ.

Output: Document Classification

For each document'd' in D do

Compute document weight in the D.

Wt=log(N/tf)*itf;

D(i)=<Docid,Index,tf,itf,Wt>;

Done

For each document id inD_i(1…n), do

TextFeatures[]=TextToWord(id);// Tokenize document with space delimiters.

Remove nonstopwords, -functional characters, special characters.

List the bio-medical features 'θ' in the TextFeaturesusing webservicekeyterms.

for each feature f in θ do

Find synonyms syn[f][] to f using GENETAG DB.

done

It is the measure of independence of thebiomedicalfeatures F;synonyms set S and the class c_m.

In our ensemble biomedical feature selection measures, enhanced chi-square and enhanced correlation based measures are used to reduce the high dimensional feature space.

$$\text{EnhancedCHISquare(ECHIS)} = \frac{P(F_i / S)[\max\{\text{Pro}(F_i, c_m)\}.\max\{\text{Pro}(F_i, \overline{c_m})\}]}{\text{Pro}(S)\,x\text{Pro}(c_m)} --(1)$$

$$\text{EnhancedCOR Relation(ECORR)} = \frac{\text{Corr}(D(F_i, j), S)[\max\{\text{Pro}(F_i, c_m)\}.]}{\text{Pro}(S)\,x\text{Pro}(c_m)} --(2)$$

Where $\text{Pro}(F_i / S)$ is the probability of the feature vector F_i that appear in the c_m and $\text{Corr}(D(F_i, j), S)$ is the enhanced correlation ofjth document features with synonym sets S.

FS[]=ECHIS; // add all feature chisquare feature measures to Feature-Set FS

FS[]=ECORR; // add all feature Correlation feature measures to Feature-Set FS

If(FS(i]>λ)

Add attribute(FSM[i], A);

End if

Else

Remove attribute (A[i],D);

End for

Done

// **Tree construction**

if(|D|==null)

MapTreeNode=null;

Create empty tree with node label null;

Else

Let $D_1, D_2, D_3 ... D_n$ be the data instances in the mapper node.

For each data instances D_i in the mapper node M do

Build the modified multinominal naïve Bayesian tree to each D_i.

To each attribute A_i in the D_i ,

compute the attribute selection measure using the equation (2).

Let A_{max} be the attribute with the maximum attribute selection measure.

Construct the MNB tree using the maximum attribute A_{max}.

Return the Mapper M tree to the Reducer phase Reduce< M_i ,return MNB(D_i)>.

Done

Done

Reducer<Mapper, Patterns>

For each Mapper in the Hadoop nodes do

Find the top k document patterns in the pattern list.

Done

4. EXPERIMENTAL RESULTS

In this experimental study, we have implementeda novel multi-class ensemble model using theHadoop framework on different medical datasets such as Medline and PubMed databases [10].We developed a textHadoop framework with the Amazon cloud server. The configuration of the Amazon AWS Server contains 10-50 cluster nodes with 10 CPU cores and 24 GB RAM to each mapper node. Amazon web services EC2, has three different types of storage sizes such as small, medium, large and this type include all kinds of cloud resources. These instance types are available for several zones. Nowadays, cloud computing technology has become

widespread in many application domains. The largestnumber of gene or protein MESH terms is used to predict the best document patterns from millions of records. Finally, we implemented and tested our hybrid ensemble classifier on biomedical disease documents to find the top patterns using gene/protein terms.

Evaluation metrics are used for calculating the performance and accuracy of machine learning algorithms. Accuracy and error rate of the classifier are used as standard metrics,however; these are not suitable to handlethe class imbalance problem as the overall accuracy issubjected to the majority class rather than a minority class with fewer samples which leads to the deprived performance of the classifier. For the class problem, general metrics are derived from a confusion matrix is shown in the below Table 1.

Table 1: Confusion Matrix

Actual_Class(AC)	Predicted_Class(PC)		
		Positive	Negative
	Positive	True_ Positive(TP)	False_ Negative(FN)
	Negative	False Positive(FP)	True Negative(TN)

The evaluation metrics associated to majority and minority classes are: 1) Recall also known as sensitivity 2) Specificity 3)Precision 4) F-measure 5)Receiver Operating Characteristic curve (ROC), 6) Geometric Mean (g-mean). Sensitivity and specificity are especially used to monitor the binary classification performance on each individual class. While precision is interested on the measurement of performance on only one class, F-measure and G-mean are used to measure the performance on both majority and minority classes.

$$\text{Recall}=\frac{True\_Positive(TP)}{True\_Positive(TP)+False\_Negative(FN)}$$

$$\text{Specificity}=\frac{True\_Negative(TN)}{True\_Negative(TN)+False\_Positive(FP)}$$

$$\text{Precision}=\frac{True\_Positive(TP)}{True\_Positive(TP)+False\_Positive(FP)}$$

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

$$GM(\#\text{Mapper}) = \sqrt{(TP/(TP+FN)).(TN/(FP+TN))}$$

ROC and F-measure are the commonly used performance metrics in classification models. However, due to noise and imbalanced problems, traditional receiver operating characteristics (ROC) and F-measure may not be a good choice. In this ensemble model, a novel phase wise accuracy measures such as geometric mean (GM) and Sum True positive rate are used to measure the performance of the classifier of the proposed model to the traditional ensemble models.

1) $GM(\#Mapper) = \sqrt{(TP/(TP+FN)).(TN/(FP+TN))}$

where TP: True Positive rate of each mapper node.

TN: True Negative rate of each mapper node.

FN: False Negative rate of each mapper node

FP: False Positive rate of each mapper node.

2) $STPR = \sum M(TPR)/N$

Where TPR is the true positive rate of each mapper,N is the total no of mapper nodes.

Table 2: Ensemble classifier performance using Mapper GMon biomedical datasets

Algorithm	**10K(documents**	**20K**	**30K**	**40K**	**50K**
C4.5	0.689	0.657	0.654	0.632	0.629
Hierarchical Multiclass DT	0.703	0.714	0.708	0.7325	0.741
Neural Networks	0.734	0.7529	0.805	0.853	0.845
SVM-PCA	0.798	0.774	0.739	0.7	0.799
HBEC	0.876	0.862	0.895	0.907	0.938
Proposed Multi-class Ensemble Model	0.9214	0.933	0.929	0.948	0.954

From the Table 2, we can see that the Mapper's average Geometric mean (GM) scores achieved by proposed model (~94%) are higher than the traditional ensemble model on different node configurations. As the size of the documents increases, proposed model improved well as compared to traditional models.

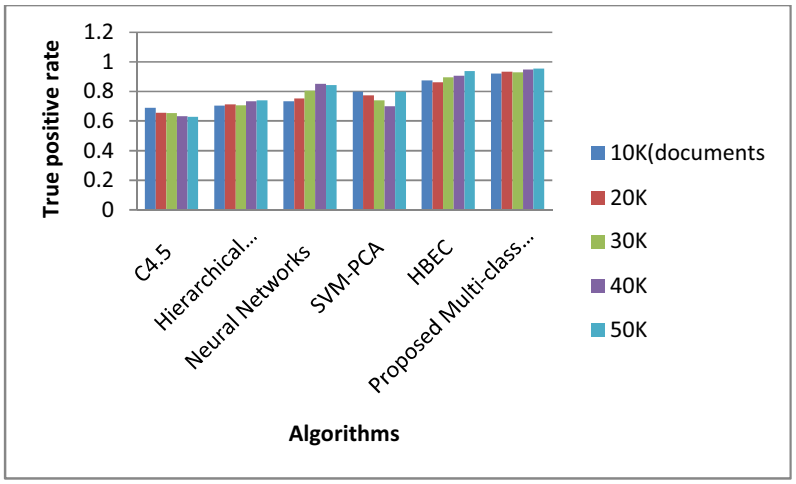

Figure 2: Ensemble classifier performance using Mapper GM on biomedical datasets

In the above Figure 2, the proposed model is compared with the different algorithms on different node mapper GM on different set of documents, it is clear that the proposed model is very efficient than other algorithms.

Table 3: Ensemble classifier performance using Mapper's STPR on biomedical datasets

Algorithm	**10K(documents)**	**20K**	**30K**	**40K**	**50K**
C4.5	0.684	0.663	0.685	0.676	0.61
Hierarchical Multiclass DT	0.732	0.787	0.779	0.787	0.708
Neural Networks	0.698	0.719	0.723	0.794	0.803
SVM-PCA	0.803	0.832	0.854	0.839	0.862
HBEC	0.857	0.872	0.881	0.917	0.957
Proposed Multi-class Ensemble Model	0.921	0.947	0.958	0.962	0.9645

From the Table 3, we can see that the Mapper's Sum True Positive Rate scores achieved by proposed model (~95%) are higher than the traditional ensemble model on different document sets. As the size of the document sets increases, proposed model improved well as compared to traditional models.

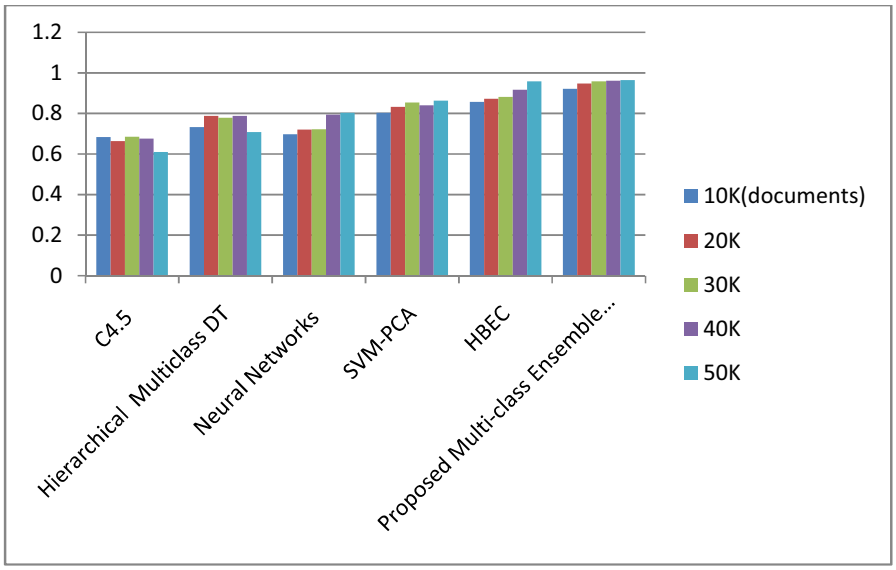

Figure 3: Ensemble classifier performance using Mapper's STPR on biomedical data

From the Figure 3, we can see that the Mapper's Sum True Positive Rate scores achieved by proposed model (~95%) are higher than the traditional ensemble model on different document sets. As the size of the document sets increases, proposed model improved well as compared to traditional models.

Table 4: Mapper/Reducer Runtime comparison of proposed model to the traditional models in Hadoop Framework

Algorithm	**MapperRuntime(ms)**	**Reducer Runtime(ms)**
C4.5	7648	5754
Hierarchical Multiclass DT	6735	4727
Neural Networks	6934	4644
SVM-PCA	7835	6467
HBEC	3647	3812
Proposed Multi-class Ensemble Model	2874	3014

From the Table 4, we can see that the Mapper's run time and Reducer runtime achieved by proposed model is less than the traditional ensemble model on document sets. Proposed model improved well as compared to traditional models.

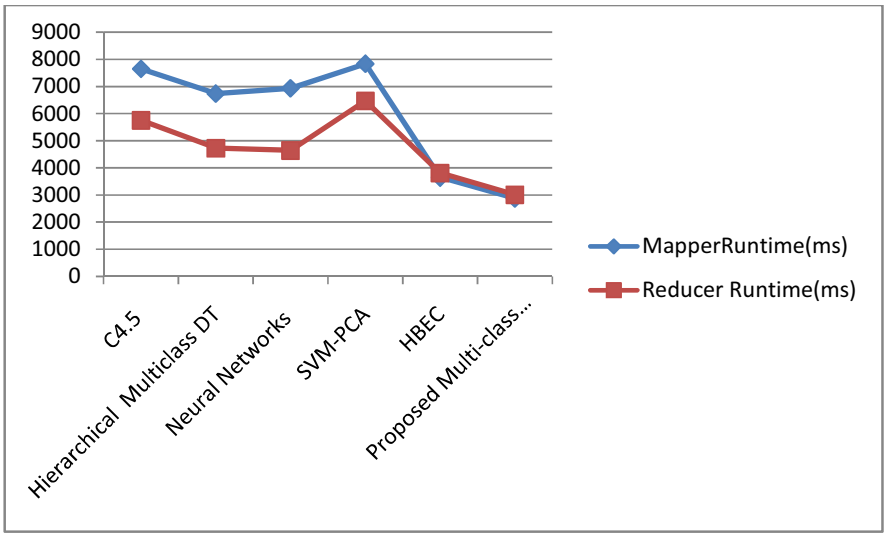

Figure 4: Runtime comparison of proposed model to the traditional models in Hadoop Framework

From the Figure 4, we can see that the Mapper's run time and Reducer runtime achieved by proposed model are less than the traditional ensemble model on document sets. Proposed model improved well as compared to traditional models.

5. CONCLUSION

In this work, a novel Hadoop based multi-class ensemble model was constructed to classify class imbalanced biomedical data. Traditional decision tree models such as multi-variate Bernoulli model, random forest and multinomial naïve Bayesian tree use attribute selection measures to decide the best split at each node of the decision tree. Basically, it is clear that data preprocessing provides better clarification than other techniques because it allows addingupdated information or deleting the irrelevant and redundant information, which helps to balance the class data.Also, the efficiency of document analysis in Hadoop framework is limited mainly due to the class imbalance problem and large candidate sets. In this paper, we proposed a two phase map-reduce framework with text preprocessor and classification model.

In the first phase, mapper based preprocessing method was designed to eliminate irrelevant features, missing values and outliers from the biomedical data. In the second phase, a Map-Reduce based multi-class ensemble decision tree model was designed and implemented in the preprocessed mapper data to improve the true positive rate and computational time. The experimental results on the complex biomedical datasets show that the performance of our proposed Hadoop based multi-class ensemble model significantly outperforms state-of-the-art baselines.

References

[1] Guyon I, Elisseeff A. An introduction to variable and feature selection.Journal of machine learning research. 2003;3(Mar):1157-82.

[2] Lusa L, Blagues R. The Class-imbalance for highdimensional class prediction.In11th International Conference on Machine Learning and Application, IEEE 2012.

[3] Liu H, Motoda H, editors. Feature extraction, construction and selection: A data mining perspective. Springer Science & Business Media; 1998 Aug 31.

[4] Satuluri N, Kuppa MR. A novel class imbalance learning using intelligent under-sampling. International Journal of Database Theory and Application. 2012 Sep;5(3):25-36.

[5] Chomboon K, Kerdprasop K, Kerdprasop N. Rare class discovery techniques for highly imbalance data. InProc.International multi conference of engineers and computer scientists 2013 (Vol. 1).

[6] Amira BR, Faouzi B, Hamid A. Medical application: Diagnosis of Alzheimer disease from MRI and documents. InModelling, Identification and Control (ICMIC), 2015 7th International Conference on 2015 Dec 18 (pp. 1-6). IEEE.

[7] Ma K, Jeong H, Rohith MV, Somanath G, Tarpine R, Schutter K, Blostein D, Istrail S, Kambhamettu C, Shatkay H. Utilizing image-based features in biomedical document classification. InImage Processing (ICIP), 2015 IEEE International Conference on 2015 Sep 27 (pp. 4451-4455).IEEE.

[8] Nguwi YY, Cho SY. An unsupervised self-organizing learning with support vector ranking for imbalanced datasets. Expert Systems with Applications. 2010 Dec 31;37(12):8303-12.

[9] Alibeigi M, Hashemi S, Hamzeh A. DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets. Data & Knowledge Engineering. 2012 Dec 31;81:67-103.

[10] Guo H, Viktor HL. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. ACM Sigkdd Explorations Newsletter. 2004 Jun 1;6(1):30-9.

[11] Zhang Y, Luo B. Parallel classifiers ensemble with hierarchical machine learning for imbalanced classes. InMachine Learning and Cybernetics, 2008 International Conference on 2008 Jul 12 (Vol. 1, pp. 94-99).IEEE.

[12] Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks. 2008 Apr 30;21(2):427-36.

[13] Li Q, Wang Y, Bryant SH. A novel method for mining highly imbalanced high-throughput screening data in PubChem.Bioinformatics. 2009 Oct 13;25(24):3310-6.

[14] Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications. 2009 Apr 30;36(3):5718-27.

[15] Kang P, Cho S. EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems. InNeural Information Processing 2006 (pp. 837-846).Springer Berlin/Heidelberg.

[16] Li H, Sun J. Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples–Evidence from the Chinese hotel industry. Tourism Management. 2012 Jun 30;33(3):622-34.

[17] Schapire RE. The boosting approach to machine learning: An overview. InNonlinear estimation and classification 2003 (pp. 149-171).Springer New York.

[18] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2010 Jan;40(1):185-97.

[19] Krzanowski WJ, Hand DJ. ROC curves for continuous data. CRC Press; 2009 May 21.

[20] Fung G, Mangasarian OL. Proximal support vector machine classifiers.InProceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining 2001 Aug 26 (pp. 77-86).ACM.

[21] GarcíaDíaz EE, Lozano Martínez F. Boosting support vector machines. Revista de Ingeniería. 2006 Nov(24):62-70.

Authors:

Thulasi.Bikku, working as Assistant Professor in Vignan's Nirula Institue of Technology and Sciences for Women. Her areas of interest are: Big Data Analytics and Security.

Dr. N.Sambasiva Rao, working as Principal in SRITW, Warangal. His areas of interest are: Software Engineering, Computer Networks and Big Data Analytics.

Dr. Akepogu.Ananda Rao, working as Director of Academics and Planning in JNTUA,Anantapur. His main areas of interest are: Software Engineering, Computer Networks, Artificial intelligence, Big Data Analytics.