THEORETICAL BIOLOGY FORUM
ISSN: 2283-7175 | https://tbfpeerjournal.org/

An Ensemble-Based Approach for Improving Classification Accuracy on Imbalanced Data

Rohini Sharma*, Piyush Gaur
* Department of Electronics and Communication Engineering, Rungta College of Engineering and Technology, Bhilai, India
Department of Electronics and Communication Engineering, Rungta College of Engineering and Technology, Bhilai, India

ABSTRACT
Classification is one of the critical task in datamining. Many classifiers exist for classification task and each have their own pros and cons. It is observed that due to imbalancing in datasets quality of classification accuracy is decreasing. Thus the increasing rate of data diversity and size decreases the performance and efficiency of classifiers. Thus it is very much important to get the maximum classification accuracy. Ensemble learning is a simple, useful and effective meta-classification methodology that combines the predictions from various classifiers. In this research an empirical study hasbeen done using voting based ensemble learning technique on varying imbalance data and varying organization for improving classification accuracy.

KEYWORDS: ensembling; classification; classification accuracy; datamining.

INTRODUCTION
Ensemble Learning basically goes with two-step decision making process, in which the first one deals with the decision of an individual classifier and second step refers to decision of combined model. The primary benefit of using ensemble systems is the reduction of variance and increase in confidence of the decision. In improvingclassification accuracy. Ensemble learning plays a vital role because of its voting ensemble learning plays a properties to get out best output for classification accuracy. The goal of classification is to assign a new entity into a class from a pre-specified set of classes [A K Saxena et al.]. Now in every sphere of an organization, data imbalancing is a major issue due to increased rate of data diversity, which slows down efficiency of classifiers and badly affects prediction. In order to overcome this issues many techniques have been invented to improve the classification accuracy of datasets. These techniques are datamining techniques which helps for making better decision and prediction. With the same purpose of learning robust models, classifier ensembling distinguishes itself by involving predictions from multiple base learners instead of relying on a single learner. By doing so, although the performance of each base learner can be relatively weak, the combined outcomes are robust for predictions. Because of its effectiveness, classifier ensembling has been applied into many research and applications.

Machine learning has many applications and is used most significantly in data mining [Sarwesh Site and Sadhana K Mishra].Data classification is the categorization of data for its most effective and efficient use. Data classification is a two-step process: 1. Learning (Model Construction) 2. Classification (Model Usage). In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or ―learning from‖ a training set made up of database tuples and their associated class labels. In the second step, the model is used for classification. A test set is used, made up of test tuples and their associated class labels. These tuples are randomly selected from the general data set [Jay Bhatt, Nikita S patel].Ensemble methods create base classifiers and the outputs are combined, usually by voting, to get better classification accuracy. Improved classification results can be achieved by using diverse classifiers [Nikita Joshi,Shweta Shrivastava.]In order to build a prediction model capable of capturing the underlying decision logics of a dataset, a straight- forward way is to improve the data quality. The data quality is determined by both intrinsic natures of the dataset and external factors.

The ensemble methodology is basically to weigh different individual classifier and combine them in order to obtain a classifier that outperforms everyone of them [A K Saxena et al.]. It has been observed that imbalance data problem occurs when negative (majority) instances outnumbers the amount of positive class instances. It is also been discussed that the decision when obtained from one given classifier may vary due to random variation in the model, thus combination of output decision of several classifiers proved an absolute method to reduce the risk of selection of poorly performed classifier. Several classification techniques are in vogue such as Logistics Regression, Classification Trees and Discriminant Analysis. Some of the fields where classification techniques find application are Engineering, Finance, and Marketing. For example, a bank would want to predict the possibility of default on part of the customer before disbursing loan to him. Similarly, a company would want to

predict the possibility of success before marketing a product in a certain area. However, one of the issues in the datasets used for prediction is that they are imbalanced. For example, in a dataset of 1000 loan disbursed, one may find 100 cases of defaults. Although, in 90% of cases in such situations there was no default, the rest 10% cases constitute tremendous loss for banks.

The main goal of ensemble method is to make a combination of various different kinds of models and each of them solves the same task to obtain the better global model. I n order to get the best results for classification accuracy by ensembling technique here various datasets from varying organization are taken into consideration for evaluation classification accuaracy . Different data mining techniques such as Logistic Regression(LR), Classification Tree (CT) and Discriminant Analysis (DA).

LITERATURE REVIEW

Classification is a data mining (machine learning)technique used to predict group membership for datainstances [D.Gopika, B.Azhagusundari]. Ensemble learning techniques are learning algorithms that generate a set of classifiers and then classify new data points by considering a (weighted) vote of their estimates. The novel ensemble technique is Bayesian averaging, however more recent techniques include error-correcting output coding, boosting and bagging [Bhavesh Patankar,Vijay Chavda].Classification of data is difficult if the data is imbalanced and classes are overlapping. In recent years, more research has started to focus on classification of imbalanced data since real world data is often skewed.

Learning and classification with imbalanced datasets has become one of the key topics in pattern recognition due to its challenges especially for real-world applications where the datasets are dominated by normal examples in addition to a small amount of unusual examples [Cigdem Beyan , RobertFisher].Data mining basically allows rethinking marketing by focusing on maximizing customer lifetime value through the evaluation of available information and customer metrics. There are several classification models, such as the Logistic Regression (LR), Classification trees (CTs) and the more recent neural networks(NNs) and support vector machines (SVMs). LR and DT have the advantage of fitting models that tend to be easily understood by humans, while also providing good predictions in classification tasks[Sergio moro et al.].Area under the curve AUC and lift proved to be good evaluation metrics. AUC does not depend on a threshold, and is therefore a better overall evaluation metric compared to accuracy. Lift is very much related to accuracy, but has the advantage of being well used in marketing practice[J Burez et al.].

VOTING BASED ENSEMBLING MECHANISM

To cope with the problem of data imbalancing and increased rate of data diversity in various organizationseveral datamining models and voting based ensemble is used to provide better classification accuracy. In this study we use three real-life application-based dataset and hence represent a sizeable test bed for comparing the datamining methods on predictiveaccuracy.The details of the procedure are shown below:

Data selection

There are three dataset were taken for this study from the sample dataset available with SPSS package(Statistical Package for Social Sciences). They are named as ‘Bankloan’,’Parole-violator’and’FlightDelays’.The dataset, named ‘bankloan’ provides details of 850 customers regarding their demographic details, debt and credit worthiness. The information on whether these customers have default or not in the past is also provided for 700 customers out of 850 in the dataset.

After removing the 150 cases where the information about default was missing, left with 700 cases on which further analysis was performed. This dataset was first divided randomly into training (70%) and test (30%) dataset, the split being on the default variable. The training dataset was used for build the model. This model was then tested on the test dataset. Each record included the output target, the default (yes, no) and candidate input features. The dataset named ‘Parole violator’ provides the details of 676 inmates regarding their demographic details of their age, male,race,state,timeserved,max.sentence,multipleoffenses,crime,violator.This dataset is randomly divided into Training(70%) and Test(30%) dataset, the split being on the violator variable. The training dataset was used for building the model. This model was then tested on the Test dataset.Each record included the output target, violator which shows the inmate have violated parole with (Yes or No). Then the third dataset named Flight Delays contains information relation to the flight to assure about whether the flight is delayed or ontime.

B.Data mining models

In this study basically three binary datamining classification techniques are used to obtain improved classification accuracy for the following dataset. The logistic regression model is used in a variety of fields:

whenever a structured model is needed to explain or predict categorical (in particular, binary) outcomes. The idea behind Logistic regression is straight forward: Instead of using Y as the dependent variable, we use a function of it,which is called a Logit .The probability of belonging to class 1(as opposed to class 0). In contrast to Y, the class number which only takes the values 0 and 1,p can take any value in the interval[0,1]. Here it is expressed p as alinear function of the q predictors in the form[$p = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta qxq$]. To use a non linear function of the predictors in the form $p = \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta qxq)}}$

In logistics regression we take two steps: the first step yields estimates of the probabilities of belonging to each class. In the binary case we get an estimate of P(Y=1), the probability of belonging to class1 (which also tells us the probability of belonging to class 0). In the next step we use a cutoff value on these probabilities in order to classify each case in one of the classes.

Classification Trees serve to limit the class of classifiers. Classification trees are easy to interpret because the representation provides a lot of intuition into what is going on. This gives confidence to the user that the classifications indeed produce the correct result. A classification tree is a model with a tree-like structure. It contains nodes and edges. The two key idea underlying classification tree are first is the idea of recursive partitioning of the space of independent variable and the second idea of pruning using validation data. The differences from regression tree growing have to do with (a) how we measure information, (b) what kind of predictions the tree makes, and (c) how we measure predictive error.

Voting Based Ensemble Algorithm

A voting based ensemble algorithm is introduced keeping in mind all the dataset and classification techniques to perform classification accuracy analysis.

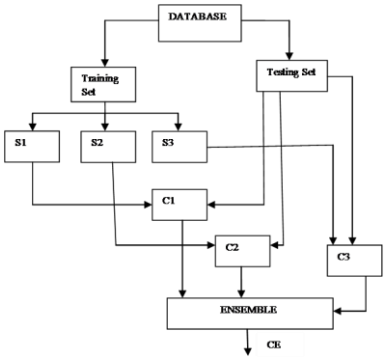

Fig1.Voting Based Ensemble Algorithm

Thus fig.1 shows how ensembling process works with different sets of data through sequence of procedure and provides required improved accuracy.Voting Based Ensemble algorithm basically deals with building a new classifier from a set of classifier for summative prediction. ensemble systems includes splitting large datasets into smaller and logical partitions, each used to train a separate classifier. This can be more efficient than using a single model to describe the entire data The primary benefit of using ensemble systems is the reduction of variance and increase in confidence of the decision.

DISCUSSION AND ANALYSIS

After studying different classification techniques combined with ensembling technique with varying imbalanced datasets has proved that classification accuracy increases as the baseline decreases . This study also shows that the among three classification techniques Classification tree provides the maximium value for the classification accuracy. It concludes that if ensemble technique is used with other datamining models, achieve improved classification accuracy and predictive performance because voting based ensemble on data mining models recovers the cons of using the data mining models alone and proved for better predictive performance.

In this project work we used R software conducted in r code for examining improved classification accuracy of an imbalanced dataset. In respect to that we used basically three classification techniques which are Logistics Regression, Classification Tree and Discriminant Analysis and along with that ensembling technique is applied to get the improved classification accuracy. In this regard we performed a comparative analysis between all the models to obtain the improved value of classification accuracy. The three dataset has been taken from SPSS(Statistical Package for Social Sciences) which is an imbalanced data named ‘Bankloan’,’Flight Delay’ and ‘Parole Violator’. Each dataset has compared and evaluated individually on the basis of the information given in

the dataset. The following observation has been obtained for all the three models for both train and test data which is shown below:

Table4: Evaluation of classification Accuracy for Test dataset

Dataset	Parameters	Baseline	Test Dataset (30%)			
			Logistic	Tree	Discriminant	Ensembling
Bankloan	TP	73.9%	33	18	32	32
	TN		140	145	145	145
	FP		15	10	10	10
	FN		22	37	23	23
	CA		82.4%	77.6%	84.3%	84.3%
	AUC		87.0%	74.9%	87.6%	
Flight Delays	TP	80.55%	9	27	9	27
	TN		532	511	532	511
	FP		0	21	0	21
	FN		119	101	119	101
	CA		82.0%	81.5%	82.0%	81.5%
	AUC		60.78%	66.17%	60.82%	
Parole Violators	TP	88.44%	0	5	2	6
	TN		178	174	176	171
	FP		1	5	3	8
	FN		23	18	21	17
	CA		88.1%	88.6%	88.1%	87.6%
	AUC		78.11%	78.11%	77.09%	

Table5: Evaluation of classification Accuracy for Training dataset

Dataset	Parameters	Baseline	Training Dataset (70%)			
			Logistic	Tree	Discriminant	Ensembling
Bankloan	TP	73.9%	64	41	52	54
	TN		333	355	341	347
	FP		29	7	21	15
	FN		64	87	76	74
	CA		81.0%	80.8%	80.2%	81.8%
	AUC		85.4%	78.6%	83.8%	
Flight Delays	TP	80.55%	23	77	23	300
	TN		1241	1206	1241	1241
	FP		0	35	0	0
	FN		277	223	277	277
	CA		82.0%	83.3%	82.0%	84.8%
	AUC		66.27%	69.91%	66.45%	
Parole Violators	TP	88.44%	2	26	2	2
	TN		413	413	411	413
	FP		5	5	7	5
	FN		53	29	53	53
	CA		87.7%	92.8%	87.3%	87.7%
	AUC		75.29%	78.11%	74.80%	

The compared accuracy between all the models ofall three dataset are discussed as follows:
For ‘Bankloan’ Dataset- The ‘Bankloan’ dataset provides extensively improved accuracy in Voting based Ensemble in training data i.e CA=81.8% and almost equal to discriminant analysis part in Test data i.e CA=84.3% . This is shown in below graph-

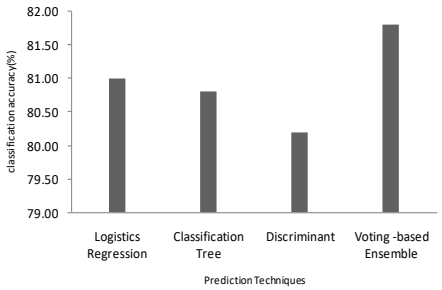

Fig 2: Classification Accuracy for Train dataset

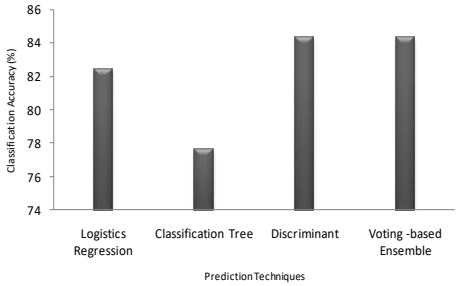

Fig 3: Classification Accuracy for Test dataset

For 'Flight Delay' dataset- In Flight Delay dataset Classification accuracy improved drastically in Voting Based Ensemble in Training data i.e CA=84.8% and lacks with logistic regression in Test data i.e CA=81.5%.

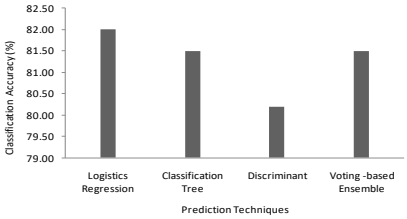

Fig 4: Classification Accuracy for Test dataset

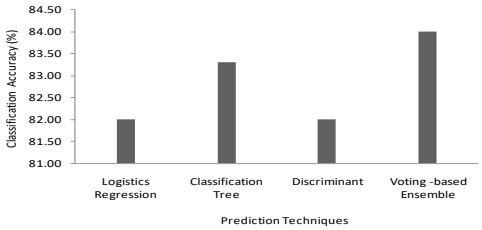

Fig 5: Classification Accuracy for Training dataset

For 'Parole-Violator' dataset-The 'Parole Violator' dataset Provides significant improvement of classification accuracyin Classification Tree models as compare to ensemble technique and other models of both Train and Test data i.e CA Train=92.8% and CA in Test= 88.6%. This is shown below

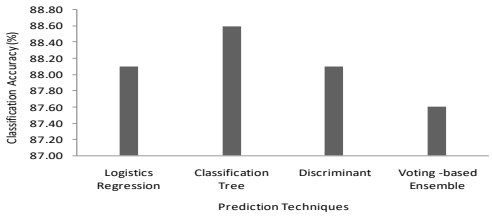

Fig6: Classification Accuracy for Test dataset

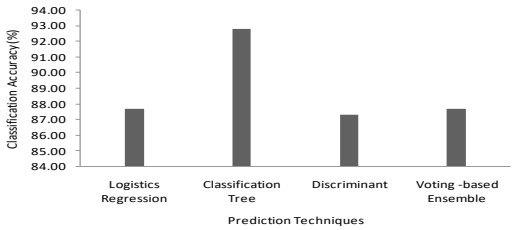

Fig 7: Classification Accuracy for Training dataset

From all the above graph it is clear that the classification accuaracy has improved significantly in training dataset part by ensembling technique and which also shows that classification tree gives a better performance than the other two models because of its high value rate of classification accuracy and we can conclude that datamining models works better as compared to statistical models. While on the test dataset part it has been observed that that the obtained classification accuracy through ensembling is almost equal to the other datamining models and here also classification Tree proves better performance due its high rate of classification accuracy.

CONCLUSION

During implementation it has been observed that voting based Ensembling technique significantly improved the classification accuracy in training dataset over the three data mining techniques logistic regression, classification tree and disriminant. It is clearly shown that the classification accuracy increases as the baseline decreases and vice versa. While on Test dataset part it has been observed that the classification accuracy obtained through ensembling is almost equal to other models in all the three dataset. Further it has viewed that the Classification

Tree proved to provide better performance due to its high classification and therefore we can conclude that Classification Tree works better for classification. Here also evaluated AUC(Area under the curve) factor which shows maximum classification done over the dataset. For bankloan dataset AUC obtained for Test dataset LR-87%,CT-74.9%,DA-87.6% and for Train dataset LR-85.4%,CT-78.6%,DA-83.8%. For Flight Delay dataset AUC obtained for Test dataset LR-60.78%,CT-66.17% AND DA-60.82% and for Train dataset LR-66.27%,CT-69.91% and DA-66.45%. For Parole Violator dataset AUC obtained for Test dataset LR-78.11,CT-78.11 and DA-77.09 and for Train dataset LR-75.29,CT-78.11 and DA-74.8. This study concludes that Voting Based Ensembling technique works proved better for providing better performance in classification accuracy. This work can be enhance using Bagging and boosting technique for future work and give assurance to have more improved classification accuracy for different organizations.

REFERENCES

[1] A K Saxena, Apoorva Mondal, Itfaq Ahmad Mir, "Improving the Classification Accuracy with Ensemble of Classifiers,"International Journal of Emerging Technology and Advanced Engineering, Volume 3, Special Issue 2, January 2013.
[2] J.Burez,D. Van DEN Poel, "Handling class imbalance in customer churn prediction," ELSEVIERJournal on Expert System with Applications, 36, 2009,pp. 4626-4636.
[3] D.J. Hand, C. Anagnostopoulos, "When is the area under the receiver operating characteristics curve an appropriate measure of classifier performance," ELSEVIERJournal on Pattern Recognition Letters, 34, 2013, pp. 492–495
[4] Sarwesh Site, Sadhana K Mishra, "International Journal of Advanced Research in Computer Science and Software Engineering," Volume 3, Issue 1, January 2013.
[5] Jay Bhatt and Nikita S patel,"A survey on one class classification using Ensemble method,"International Journal for Innovative Research in Science & Technology,Volume 1, Issue 7 ,December 2014.
[6] Cigdem Beyann, Robert Fisher "Classifying imbalanced data sets using similarity based hierarchical decomposition" ELSEVIERJournal on Pattern Recognition Letters ,48, 2015, pp.1653–1672.
[7] Nenad Tomasev and Dunja Mladenic"Class imbalance and the curse of minority hubs", ELSEVIERJournal on Knowlwdge Based System , 53, 2013, pp.157–172.
[8] C.Ferri,J.Hernandez, O.R.Modroiu, "An experimental comparison of performance for classification," ELSEVIERJournal on Pattern Recognition Letters, 30, 2009, pp. 27-38.
[9] V.Garcia,J.S.Sanchez,R.A.Molline"Onthe effectiveness of pre-processing methods when dealing with different levels of class imbalance," ELSEVIER Journal on Knowledge-Base Systems, 25, 2012, pp. 13-21
[10] Che-Chang Hsu, Kuo-Shong Wang, Shih-Hsing Chang, "Bayesian decision theory for support vector machines: Imbalance measurement and feature optimization," ELSEVIERJournal on Expert system with Application, 38, 2011,pp. 4698-4704.
[11] Thomas Verbraken, Cristian Bravo,Richard Weber, Bart Baesens, "Development and application of consumer credit scoring models using profit-based classification measures,ELSEVIER Journal on European Journal of Operational Research, 238, 2014, pp. 505-513.
[12] Nikita Joshi, Shweta Srivastava "Improving Classification Accuracy Using Ensemble Learning Technique Using Different Decision Trees" International Journal of Computer Science and Mobile Computing Vol.3 Issue.5, May- 2014, pp. 727-732.
[13] Wen-Chin,Chiun-Chieh Hsu,Jing-Ning Hsun "Adjusting and generalizing CBA algorithm to handling class imbalance" ELSEVIERJournal on Expert system with Application 39,2012,pp. 5907-5919.
[14] Chris Seiffert,Taghi M. Khoshgoftaar,Jason Van Hulse,Andres Folleco "An empirical study of the classification performance of learners on imbalanced and noisy software quality data" ELSEVIERJournal on Information Sciences 259,2014, pp.571-595.
[15] Loris Nanni, Carlo Fantozzi, Nicola Lazzarini "Coupling different methods for overcoming the class imbalance problem" ELSEVIERJournal on Neurocomputing 158,2015,pp.48-61.
[16] Myoung-Jong Kim, Dae-Ki Kang,Hong Bae Kim, "Geometric mean based boosting algorithm with over sampling to resolve data imbalance problem for bankruptcy prediction" ELSEVIERJournal on Expert system with Application 42,2015,pp. 1074-1082.
[17] Improving classifier performance using feature selection with Ensemble learning. "International journal of scientific Research in Computer Science,Engineering and Information Technology" volume1,Issue1, july 2016.
[18] D.Gopika, B.Azhagusundari, "An Analysis on Ensemble methods in classification Tasks"International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 3, Issue 7, July 2014.

