SOFT COMPUTING FOR POST OPERATIVE HUMAN LIFE EXPECTANCY IN THORACIC SURGERY

[1]Hiren Bhalala, [2]Maulin Joshi

Student, M.E.-E.C., GTU, Gujarat, India. Professor, EC Department, SCET, Surat, India.

Email:[1]hhbhalala.708@gmail.com, [2]maulin.joshi@scet.ac.in

ABSTRACT

Post operative human life expectancy is very critical issue in the medical decision problem. Skewness in the data is major problem in thoracic surgery dataset. We present some sampling and ensemble techniques that deals with the problem of imbalanced data. The goal is to know whether patient will survive(negative examples) for one year after surgery or will die(positive examples) in the span of one year. We use some filter as preprocessing of data like Synthetic Minority Oversampling Technique (SMOTE) , Under sampling technique and Discretization of data. There is improvement in the performance of Naive Bayes , SVM and Multilayer Perceptron Classifier. We also use ensemble technique that combines two or more classifier. The model has improvement in terms of accuracy of classifier and also efficient classification of critical(positive examples) instances.

Keywords Classification, Imbalanced data-sets ,Bagging , Boosting, Ensembles, Data Mining

1. INTRODUCTION

Thoracic Surgery is repair of organs located in the thorax or chest. The thoracic cavity lies between the neck and the diaphragm. Thoracic surgery repairs diseased or injured organs and tissues in the thoracic cavity. General thoracic surgery deals specifically with disorders of the lungs.

For the patient of lung cancer it may happen that the patient will die in the short time period after the surgery. In that case the surgery is not the good option. Our goal is to make a model that can accurately predict that after the surgery, patient will live for 1 year or die in the 1 year.

2. IMBALANCED DATA CLASSIFICATION TECHNIQUES

Imbalanced data problem is widely observed in many clinical decision problem. Classifiers give results biased towards majority class.

2.1 Classification

Classification means the identification of the class or category of an object. We have the previously defined classes. We want to create an automatic classifier that can classify objects that it has not yet seen but they are expected to belong in one of the classes for which the classifier is built. Classifier has to remember the previously classified examples. Classifier is build on the training data. We can select the training set and training method according to our requirement. We need to know the correct classes called labels in the training set. We need algorithms that learn from training set and also remember. Testing and comparing algorithms is done using a test set, for which the labels are known. Many algorithms also use a validation set to manage its learning process. n-fold cross validation seperates the given dataset into n different sets. (n-1) sets are used for training and 1 for testing. Thus n different classifiers are built and average of all is taken as the result.

2.2 Performance Measures

In case of imbalanced datasets the prediction done by any classifier is biased towards the majority class. Accuracy of the classifier will always be high even if all minority examples are classified as majority because of imbalanced data. In case of Thoracic Surgery patient decision mistake made in the case of a patient who will die in 1 year(positive examples) is more

troublesome than decision mistake made in the opposite direction(negative examples). Thus correctly classified positive examples defines the efficiency of prediction. We find Confusion Matrix, True Positive rate (TP_{rate}) , True Negative Rate (TN_{rate}), Geometric Mean (G_{mean}) and Harmonic Mean (F-measure).

Table 1 Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

TP-Actual positive and classified as positive
TN-Actual Negative and classified as negative
FP-Actual negative but classified as positve
FN-Actual positive but classified as negative

$$TN_{rate} = \frac{TN}{TN+FP} \quad (1)$$

$$TP_{rate} = \frac{TP}{TP+FN} \quad (2)$$

$$Gmean = \sqrt{TP_{rate}\, TN_{rate}} \quad (3)$$

$$F - measure = \frac{2*TP_{rate}*TN_{rate}}{TP_{rate}+TN_{rate}} \quad (4)$$

Gmean is the most commonly used quality rate in case of imbalanced data.

2.3 Parameters for Thoracic Surgery Dataset

The datset is taken from the UCI machine repository. It includes data of the patients from 2007 to 2011. The data is of lung cancer patients.

DGN: Diagnosis
specific combination of ICD-10 codes for primary and secondary as well multiple tumour if any. (DGN1, DGN2, DGN3, DGN4, DGN5,DGN6, DGN8.)

PRE4: Forced Vital Capacity
The amount of air which can be forcibly exhaled from the lungs after taking the deepest breath possible. (Min-1.4, max-6.3, mean-3.3, S.D.-0.9)

PRE5: Forced Expiratory Volume(FEV1)
The amount of air which can be forcibly exhaled from the lungs in the first second of a forced exhalation. (Min-0.96, Max-86.3, Mean-4.6, S.D-11.8)

AGE:
age at surgery(Min-21, Max-87, Mean-52.5, S.D.-8.7)

PRE6: Performance status
Zubrod scale (PRZ2, PRZ1, and PRZ0).
There are different ways of assessing general health. The World Health Organization designed the scale that doctors use most often.
It has categories from 0 to 5

0 you are fully active and more or less as you were before your illness
1 you cannot carry out heavy physical work, but can do anything else
2 you are up and about more than half of the day and can look after yourself, but are not well enough to work
3 you are in bed or sitting in a chair for more than half of the day and you need some help in looking after yourself
4 you are in bed or a chair all the time and need a lot of looking after yourself
5 death

PRE7: Pain before surgery-T(true) or F(false)

PRE8: Haemoptysis
Coughing up blood-T(true) or F(false)

PRE9: Dyspnoea
Difficulty in breathing-T(true) or F(false)

PRE10: Cough before surgery-T(true) or F(false)

PRE11: Weakness before surgery-T(true) or F(false)

PRE14: Size of tumour
T in clinical TNM - size of the original tumor, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13)

PRE17: Diabetes-T(true) or F(false)

PRE19: MI upto 6 months-T(true) or F(false)

PRE25: PAD(Peripheral arterial diseases)-T(true) or F(false)

PRE30: Smoking-T(true) or F(false)

PRE32: Asthama -T(true) or F(false)

RISK1Y: 1Year survival period-T(true) or F(false)
True value if died

2.4 Preprocessing

In this approach before applying any classifier,we do preprocessing on the dataset. In preprocessing we can apply a filter on the data. There are two types of filter (1) supervised filter and (2) non supervised filter. In each type of filter we can balance the attribute and instances also. For example, we can give some range to the given attributes and replace the value with intervals. This filter is applied on the attribute. Similarly we can increase or decrease the number of insatances in each of the class.

2.4.1 Undersampling

Every single classifier give the result biased towards the majority class. This problem of large instances can be solved by elliminating some majority examples ramdomely. This is called undersampling. We can call it supervised filter applied in the instances. After applying that filter the number of positive and negative examples will be same as that of positive instances.

2.4.2 Oversampling

The method to increase the number of positive examples is called oversampling. SMOTE is a technique that increases positive instances. This method put a synthetic example between two positive examples. This new example is also labeled as positive example with assumption. Sometimes it may happen that one negative example is there between two positive, then putting synthetic sample between them is a bad idea. This can be overcome by putting synthetic sample on the line joining k- nearest negative neighbours of a positive example. The number of synthetic samples depends upon the number of nearest neighbor selected.

2.4.3 Discreatization

There are different types of attributes such as nominal, numeric, binary etc. Suppose we have a training data in which age of a person is from 21 year to 85 year and in testing set , there is a person with age 20 year. Then the system will be confused to classify that instance. In order to deal with this problem, we have given the range instead of the value. This technique is called discreatization. This filter is applied on the attributes and is unsupervised filter.

2.4.4 Remove attributes

In every dataset, there is some attributes which affects the output the most.To verify the effects of different attributes, we can remove some attributes and check the efficiency of the classifier. By this we can also find the relative gain of the attributes. We can extract rules based on the information gain of particular attribute.

2.4.5 Remove Folds

n-fold cross validation creates n datasets from the given dataset. (n-1) fols are used for training and 1 for testing. Similarly another (n-1) folds are used for training and 1 for testing and so on. At the end n different model's average result is taken. If we want to check on which fold the classifier perform the best then we can remove other folds and extract information about the subset of the whole dataset.

We can also combine two or more technique to improve the performance.

2.5 Algorithmic approach

In the algorithmic based approach, continous correction and improvement is done in the training process. Different classifier works differently on the different data depending upon the nature of the data. That different classifier can be run on the data on which they perform better. After that all the classifier are combined to build the overall model. This method of combining the classifier is called ensemble learning.

2.5.1 Ensemble learning

There is no algorithm that is always the most accurate.The goal is to Generate a group of base-learners which when combined have higher accuracy. We can combine the classifiers by different strategy.

1. Average of probability
2. Product of probability
3. Maximum probability
4. Minimum probability
5. Majority voting
6. Median

In bagging we can take m bootstrap sample and train a classifier on the different samples. Then combine the classifier.It is called bootstrap aggregating(Bagging). But in boosting every classifier is given some weights according to the classifier's performance. Weighted sum is taken to add the classifiers.

3. EXPERIMENTAL RESULTS

Preprocessing is done before we apply the imbalanced dataset to a classifier to balance the data. Results of different preprocessing technique and combination of that techniques are obtained. Different classifier results are observed and the best results are extracted by tuning the parameter. Accuracy of classifier means correctly classified instances is important but the most important thing is to have higher true positive rate and higher geometric mean of true positive rate and true negative rate.

3.1 Classifiers used in the experiments

We have taken different classifiers to check the performace of the model. Some of them which has good result is listed below.

- SMO
- Naive Bayes
- Multilayer Perceptron

- LibSVM(C-SVC)
- Vote

SMO stands for Sequencial Minimal Optimization is a support vector classifier. It generally used for sparse dataset. LibSVM is a library for svm and train a cost sensitive support vector machine. Naive Bayes is the baysian classifier work under the bayes' rulr for postpriory probability. Multilayer Perceptron is a n-layered feed forward network worked on the back propagation algorithm.

3.2 Preprocessing

For the problem of imbalanced data,we have used over sampling, undersampling, descretization and also combine two or more. This will balances the data and improve the predictive accuracy and other performance parameters.

- SMOTE
  - Put synthetic minority sample between k-nearest positive sample to increase the minority instances.
- Spread Sub Sample
  - Randomly elliminate the negative sample to balance the data
- Discretization
  - Give range to the numeric attribute to generalize the model
- Combination of above

3.3 Test Options

After training we can have different test options We have used

- Use Training Set
- 10-fold Cross Validation
- 3-fold Cross Validation

3.4 Performance Parameters

In case of Imbalanced Dataset, True positive rate and Geometric mean are the impotant parameters. We have calculate and compared the parameters given below.

- TPrate-True Positive Rate
- Gmean-Geometric Mean
- Accuracy of Classifier

3.5 Comparison of different pre-processing and classifier

We use different pre-processing technique and different classifier on the thoracic surgery dataset. Analysis is done in terms of Accuracy of the classifier, True Positive Rate and Geometric Mean of True positive rate and true negative rate.

3.5.1 Without Preprocessing

Without pre-processing, classifier decision is biased towards the majority class. Almost all the positive examples are classified as negative. Thought the accuracy of the classifier is about 85% because of imbalanced data, the performance is considered very poor. This is because Correctly classified positive examples are nearly equal to zero.

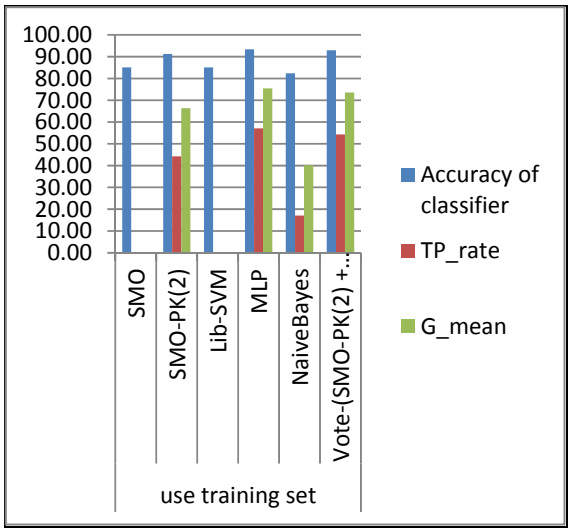

Figure 3.1 Performance parameters without preprocessing(Use training set)

Figure 3.1 shows the result for the different classifier applied without pre-processing. Test option is use training set means whole training set is used for testing. Accuracy of the classifier is defined as the correctly classified instances. Here Vote is an ensemble classifier which is used to combine more than one classifier. The result is for the ensemble of SMO with polynomial kernel of degree 2 and Multilayer perceptron.

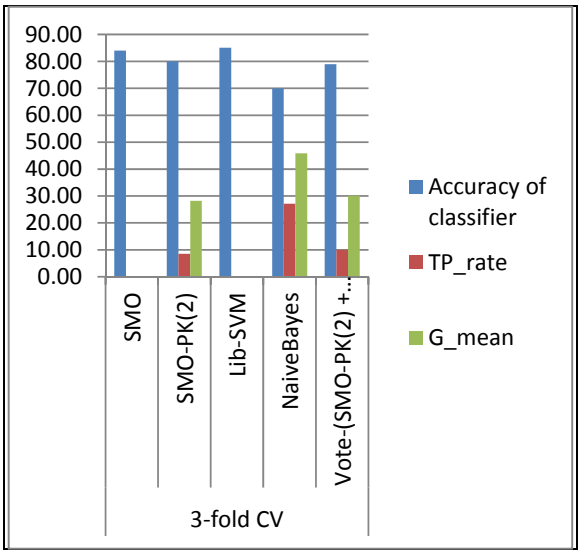

Figure 3.2 Performance parameters without preprocessing (3 fold CV)

Figure 3.2 shows the results for the test option 3-fold cross validation. It means whole training set is divided into 3 folds. Two folds are used for training and one for testing and one by one taking all folds for testing. The average of all the models is taken as final result.

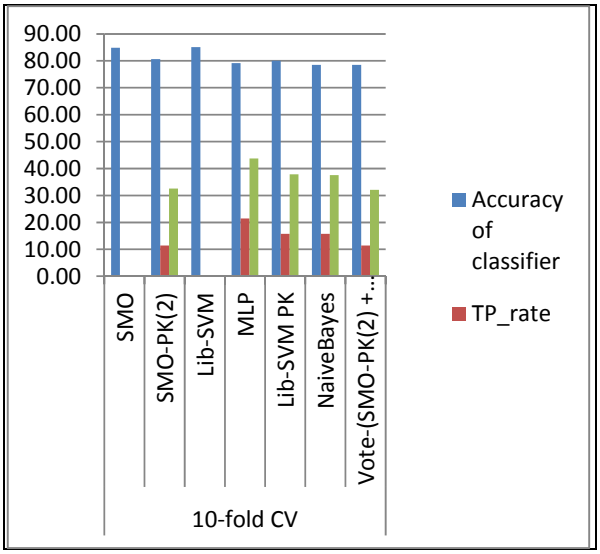

Figure 3.3 Performance parameters without preprocessing (10 fold CV)

Figure 3.3 shows the result with test option 10-fold cross validation similar to 3-fold cross validation. True Positive rate is defined as

$$TP_{rate} = \frac{TP}{TP + FN}$$

It represents the correctly classified positive examples. For comparison purpose in graph it is multiplied by 100.

3.5.2 Preprocessing-SMOTE

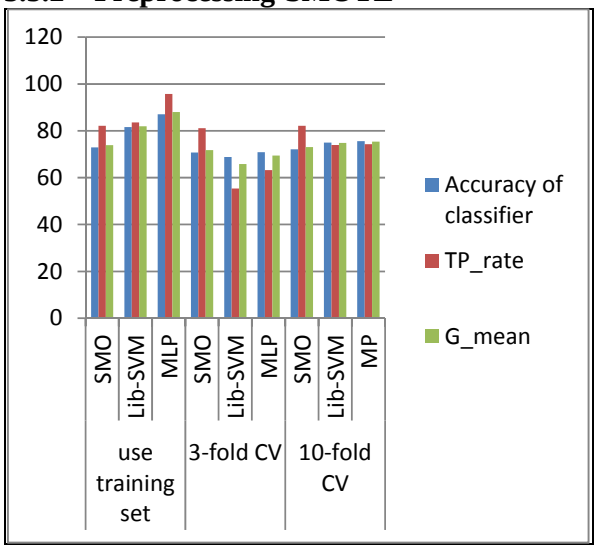

Figure 3.4 Performance parameters with SMOTE

Figure 3.4 shows the classifier performance with SMOTE. It is Synthetic Minority oversampling technique used to increase the number of minority examples.G_mean is geometric mean of TPrate and TNrate. Multilayer perceptron gives better geometric ratio.

3.5.3 Preprocessing-Spread Sub Sample

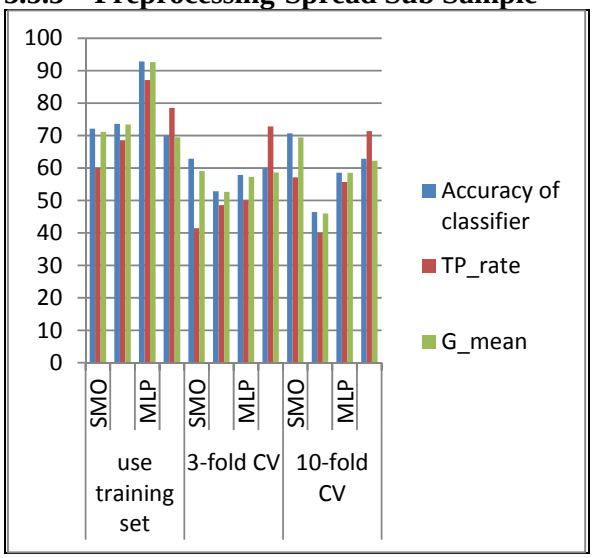

Figure 3.5 Performance parameters with Spread Sub Sample

Figure 3.5 shows the result with Spread Sub Sample which is the under sampling technique. Its reduces the majority samples to the minority sample and hence balances the classes to improve classifier accuracy. Lib-SVM is a library for support vectors and used to classify with SVM classifier.

3.5.4 Preprocessing- Discretization

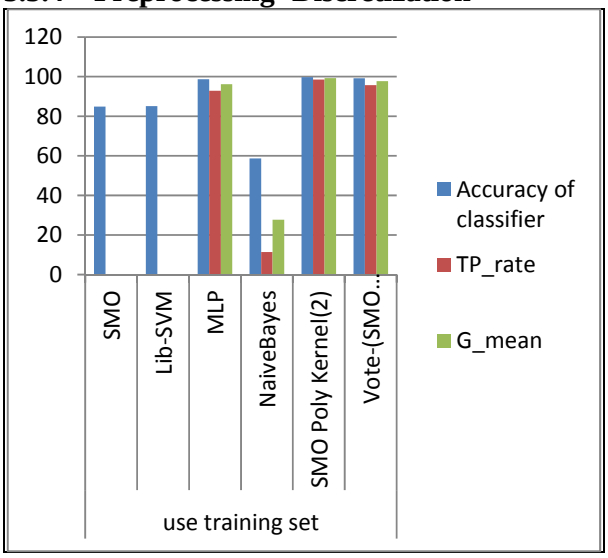

Figure 3.6 Performance parameters with Discretization(use training set)

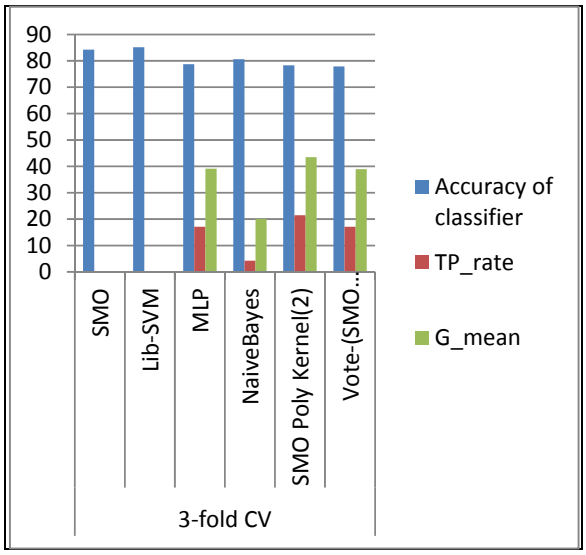

Figure 3.7 Performance parameters with Discretization(3- fold CV)

Figure 3.6, 3.7, 3.8 shows the accuracy of different classifier with Dicretization. Disretization gives range to the attributes that are numeric. It generalise the model. Figure shows the results for test options use training set, 3- fold cross validation and 10- fold cross validation respectively.

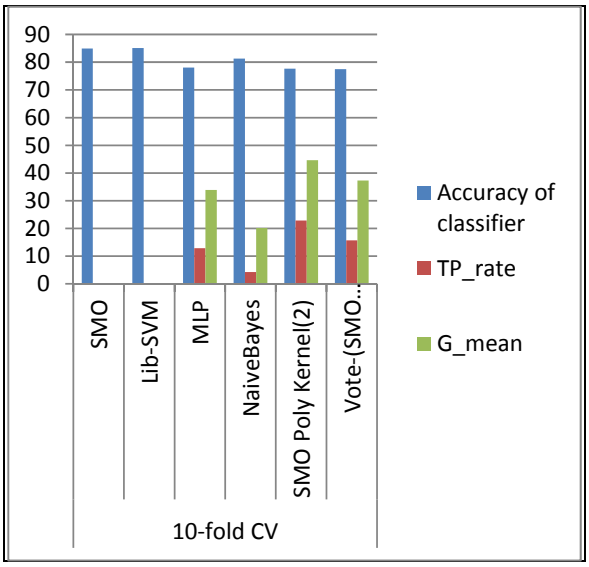

Figure 3.8 Performance parameters with Discretization(10- fold CV)

3.5.5 Preprocessing- Discretization+SMOTE

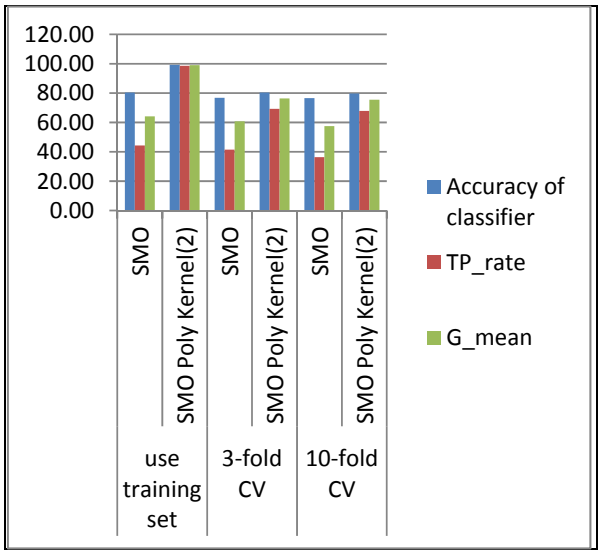

Figure 3.8 Performance parameters with (Discretization+SMOTE)

Figure 3.9 shows result for the combination of two pre-processing techniques Discretization and SMOTE. SMO Poly Kernel uses polynomial kernel of degree two which transform the data to higher dimensional space and thus improves performance.

3.5.6 Preprocessing-Discretization+ SMOTE+ SpreadSubSample

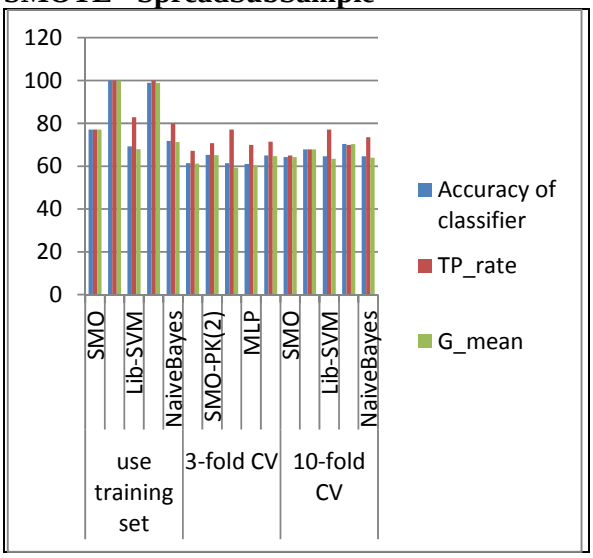

Figure 3.9 Performance parameters with (Discretization+SMOTE+SpreadSubSample)

Figure 5.10 shows the accuracy for combination of Discretization, SMOTE and Spread Sub Sample. Naive Bayes classifier is a Bayesian classifier that uses bayes formulae to classify with strong assumptions of independency.

3.5.7 Preprocessing-SMOTE+SpreadSubSample

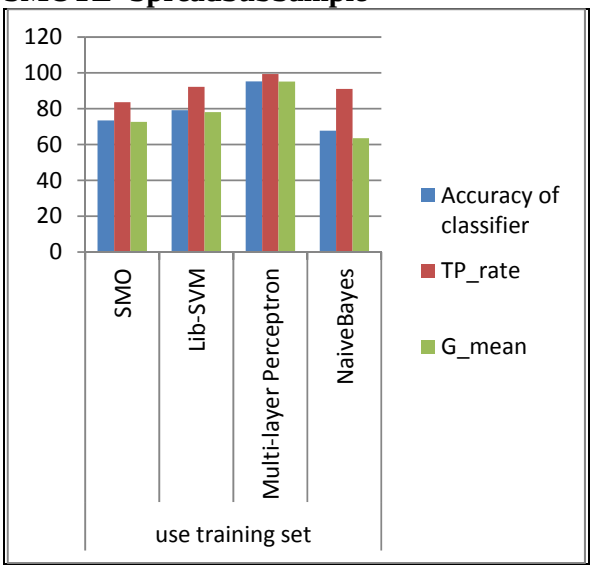

Figure 3.10 Performance parameters with SMOTE+SpreadSubSample(use training set)

Figure 3.11,3.12 shows results for combination of SMOTE and Spread sub Sample for test option use training set and 3-fold CV. It is the combination of under sampling and over sampling. It increases the minority examples and decreases the majority examples and hence combines the advantage of both the techniques.

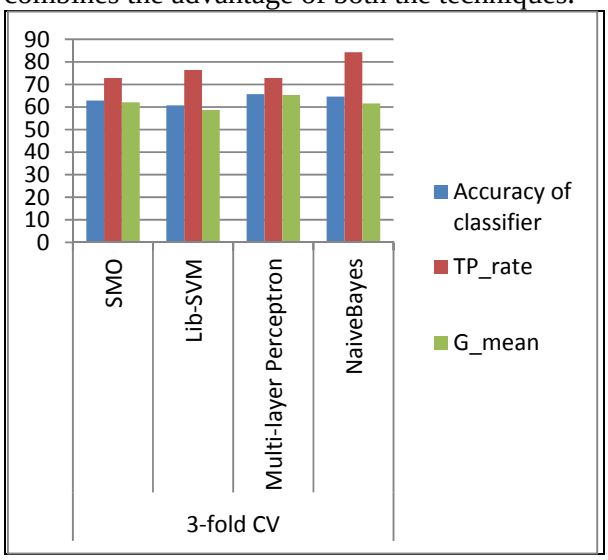

Figure 3.11 Performance parameters with SMOTE+SpreadSubSample(3-fold CV)

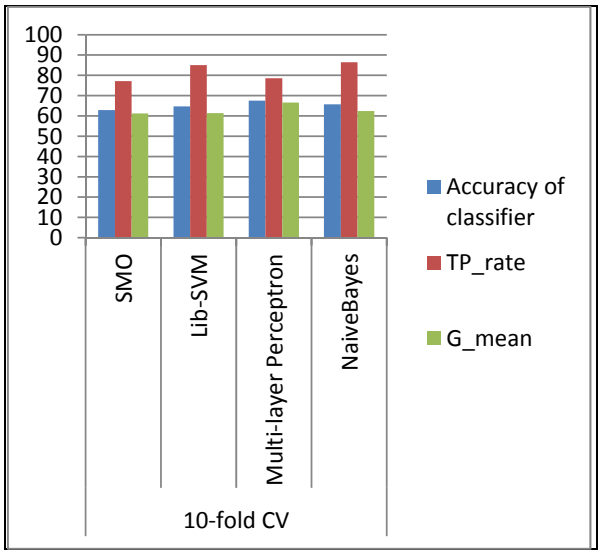

Figure 3.12 Performance parameters with SMOTE+SpreadSubSample(10-fold CV)

Figure 3.13 shows results for combination of SMOTE and Spread Sub Sample with test option 10-fold cross validation. Multi Layer Perceptron and Naive bayes performs better in terms of correctly classified positive examples.

4 CONCLUSION

Single classifier gives prediction biased towards the majority class. Preprocessing techniques like descretization, undersampling and over sampling improves the true positive rate and geometric mean along with predictive accuracy of the

classifier. Multilayer perceptron works better than naive bayes classifier and SMO with linear kernal. Disretization with SMO with polynomial kernal improves true positive rate

5 REFERENCES

[1] Zia ba, M., Tomczak, J. M., Lubicz, M., & AswiA tek, J. "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients."Elsevier, 2014

[2] Galar, M.; Fernández, A; Barrenechea, E.; Bustince, H.; Herrera, F., "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", *IEEE Transactions on Systems, Man, and Cybernetics*, Part C: Applications and Reviews, vol.42, no.4, pp.463,484, July 2012

[3] Chorowski, J.; Zurada, J.M., "Extracting Rules From Neural Networks as Decision Diagrams," *IEEE Transactions on Neural Networks*, vol.22, no.12, pp.2435,2446, Dec. 2011.

[4] Liu, Y., An, A., & Huang, X. "Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles" *Springer, vol. 3918*, pp.107-118,2006.

[5] Mohammed J. Zaki , "DATA MINING TECHNIQUES" Department of Computer Science, Rensselaer Polytechnic Institute Troy, New York , USA,August 9,2003

[6] Zoya Gavrilov, "SVM Tutorial", Massachusetts Institute of Technology(available online PDF(http://web.mit.edu/zoya/www/SVM .pdf))

[7] Roy, K., C. Chaudhuri, M. Kundu, M. Nasipuri, and D. K. Basu. "Short Paper_."*Journal of information science and engineering*"21 (2005): 1247-1259.

[8] Caruana, R. and Niculescu-Mizil, A.: "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning, 2006.

[9] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-SMOTE: a new over-samplingmethod in imbalanced data sets learning, in: D.-S. Huang, X.-P. Zhang, G.-B.Huang (Eds.), Advances in Intelligent Computing, 2005, pp. 878–887

[10] C. Seiffert, T. Khoshgoftaar, J. Van Hulse, A. Napolitano, Rusboost A hybridapproach to alleviating class imbalance, IEEE Transactions on Systems, Manand Cybernetics, Part A: Systems and Humans 40 (2010) 185–197.

[11] N. Chawla, A. Lazarevic, L. Hall, K. Bowyer, Smoteboost: improving prediction of the minority class in boosting, in: N. Lavrac, D. Gamberger, H. Blockeel, L.Todorovski (Eds.), Knowledge Discovery in Databases: PKDD 2003, Springer,Berlin Heidelberg, 2003, pp. 107–119.

[12] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data