DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5523

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PERFORMANCE ANALYSIS OF PREDICTING SURVIVAL RATES IN IMBALANCED HEALTHCARE DATASET

R.Vani
Funmeld Networks
Chennai, India

Abstract: Predicting Patients health is a critical task in the Healthcare Industry. Healthcare datasets show a high degree of imbalance especially for rare diseases. The current work aims at predicting the post operative survival rate in thoracic surgery datasets. The dataset exhibits data imbalance with around 15% positive cases and remaining 85% negative cases. The commonly applicable machine learning techniques for prediction score poorly in predicting the positive cases in spite of high accuracy of the predictions for the negative cases. We use SMOTE (synthetic minority oversampling technique) to reduce the degree of imbalance and increase the positive samples proportion before the application of the following classifiers: Naive Bayes, Neural Networks, Random Forest, Boosting algorithms - Adaboost, Extreme Gradient boosting and Support Vector Machines and examine the results. The study shows that SVM and Naïve Bayes show significantly better performance on the imbalanced datasets than other models using synthetic datasets than under normal conditions.

Keywords: Classification, healthcare, Thoracic, SMOTE, Synthetic datasets, lung cancer, SVM, Naive Bayes.

I. INTRODUCTION

Recent advancements in healthcare have caused majority of practitioners to rely on machine learning based models for assessing the health of the patients improving their ability in providing diagnostics, prognosis and prediction of the effect of treatments in clinical studies. Majority of health conditions generate data that have significantly less number of positive cases than negative cases. Such situations can be seen in diagnosis of different types of cancer, diabetes, post surgery survival rates and rare diseases. In other real world scenarios, this data imbalance problem is also predominant in scenarios where anomaly detection is crucial like electricity pilferage, fraudulent transactions in banks, intrusion detection systems. The predictive model developed using conventional machine learning algorithms could be biased and inaccurate for these applications. Using standard classifiers leads to bias towards the larger class with predictions done on the smaller class being categorized as misclassifications errors. However, as most of the machine learning algorithms require that the class sizes be more or less similar in size for ensuring correct results, most of the models cannot be used for assessment in healthcare systems. Hence correction strategies addressing the data imbalance issue must be incorporated before classifiers are used on the data. Common approaches for handling these issues are data based approach and algorithm based approach.

Data Based Approach

In the data-based approach, also known as the sampling approach, sampling technique is used to overcome the problem without altering the classification algorithm. Being the most common solution, we can either perform under-sampling of the majority class or over-sampling the minority class or both. Also the SMOTE technique of synthetic data generation can be regarded as a type of oversampling technique [1]. SMOTE creates new minority class examples interpolating between several minority class examples that lie together, using the k-nearest neighbor algorithm. This ensures that the over fitting problem be avoided causing the decision boundaries belonging to minority class to spread into the majority class space.

Algorithm Level Approach

In the algorithm based approach, we modify the standard classification algorithms to rectify the imbalance. Standard classification algorithms generally use a default decision threshold to assign class membership for maximizing the classification accuracy. This is based on an assumption of an equal cost of misclassifications. Ensemble approaches also may be used to solve the problem of uneven data in the training phase.

The aim of the paper is twofold: Firstly to identify machine learning techniques which provide acceptable classifier performance through the use of select pre-processing techniques as SMOTE and secondly to determine valid metrics that can be used to measure performance of imbalanced data. The paper is organized as follows: Section 1 provides a brief introduction to the approaches in handling the data imbalance problem. In Section 2 we take a look at work done on similar healthcare datasets concerning prediction or classification types of problems. The Research Methodology used in classification of the dataset is examined in Section 3. Section 4 provides an overview of the different machine learning models that are used as classifiers in predicting the survival rate of the patients. Details regarding the dataset used are discussed in Section 5. In Section 6, we look at the performance metrics that are applicable to the current dataset considering the imbalance issue. In Section 7, we summarize our results and discuss the findings. We draw conclusions based on our findings in Section 8.

II. Related work

Increasing number of papers has been put forth on prognosis and diagnosis of Diseases with a large subset of them dealing with the imbalance issue. We take a look at a few of the papers below: Using the thoracic surgery dataset, authors in [2] use the SVM with boosting to overcome the imbalance issue. The boosted SVM is used for extraction of decision rules using an oracle based approach which can be used for the prediction of life expectancy of lung cancer patients. The current work draws the dataset from this work and improves on the performance of the Boosted SVM. In [3] the researchers discuss in their review paper, the work done in prediction of heart disease for machine learning models using Naive Bayes, Neural Networks, Decision Trees, and also takes a look at deep learning algorithm used for the above purpose. In [4] the authors in their work combine KNN with genetic algorithm for effective classification leading to enhanced prediction of Heart Disease. Paper [5] uses rough sets for the prediction of Breast Cancer and the rough sets create decision rules which are used by a MATLAB program for future diagnosis of the disease. In [6] we see the authors discussing a number of supervised learning techniques and applying them to the SEER database to classify lung cancer patients in terms of survival, including linear regression, Decision Trees, Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and a custom ensemble. Key data attributes in applying these methods include tumor grade, tumor size, gender, age, stage, and number of primaries. The prediction is treated as a continuous value instead of categorical in this case. In [7] Bergquist et al. develop tools for tools for classifying lung cancer patients receiving chemotherapy into early vs. late stage cancer using an ensemble machine learning model and creating a set of classification rules for the predicted probabilities. Schubach et al. present a novel method in [8] that adopts imbalance-aware learning strategies based on re-sampling techniques and a hyper-ensemble approach which outperforms state-of-the-art methods. Mustafa et al. [9] in their work, propose an efficient combined algorithm based on Farther Distance Based on Synthetic Minority Oversampling Technique and Principle Component Analysis. The method successfully reduces the high dimensionality and balances the minority class. .

III. Research Methodology

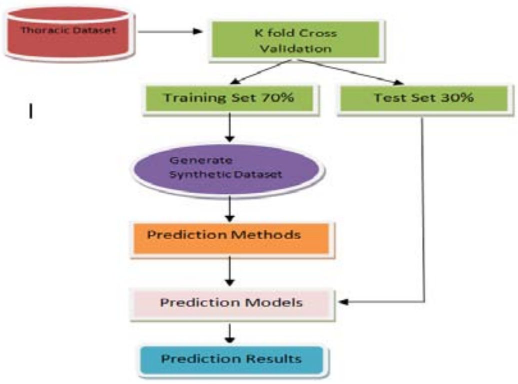

Figure 1. Research Methodology Adopted.

Prediction of Surgical Rates is of immense importance to healthcare professionals and there is an immense need to utilize the machine learning techniques to learn from past data and recognize patterns. The general framework of the Models is show in Fig.1. We take proper care to ensure that synthetic datasets are generated from the training set during the cross validation phase and keep the test set unadulterated for estimation purposes. We apply the method to each of the models chosen to get the prediction results for comparative study.

IV. Predictive machine learning classification

In this section, we take a look the various predictive machine learning classification models relevant to the current dataset.

A. *Support Vector Machines*

A Support Vector Machine (SVM) is a discriminative classifier defined in formal terms by a separating hyper plane. When labeled training data *(supervised learning)* is used, the SVM algorithm outputs an optimal hyper plane which is capable of categorizing new examples. SVM give the largest minimum distance to the training examples. Twice, this distance is termed as the Margin. Hence, the optimal separating hyper plane *maximizes* the margin of the training data [10]. Various SVM algorithms use different types of kernel functions including linear, nonlinear, polynomial, radial basis function and sigmoid.

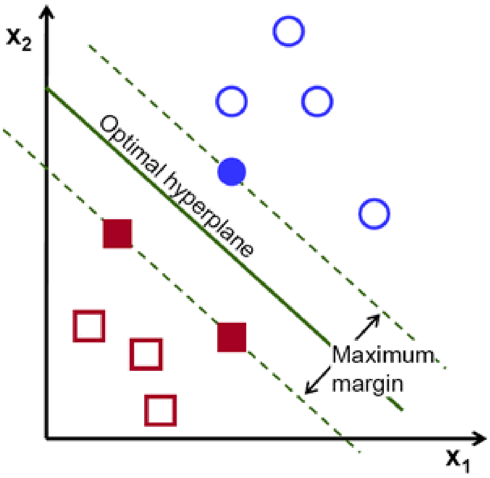

Figure 2.Support Vector Machine

B. *Naive Bayes*

Naive Bayes algorithm is based on conditional probability which takes into account the prior knowledge. The classifier predicts the membership probability for each class and assumes the effect of the value of a predictor on a given class is independent of the other predictor values. The class with the larger probability is considered as the likely class. This is termed as Maximum a Posteriori. Also the classifier makes an assumption that all the features are unrelated and hence cannot be used for learning relationship between the variables. The classifier is simple to use since it has no iterative estimation and can outperform other complex algorithms on large datasets despite its simplicity.

C. *Neural Networks*

An artificial neuron is a device with many inputs and one output. Neurons are modeled after real life biological neurons and have a similar structure as them. The neuron has two modes of operation; the Training mode and the Using mode. During the training the neurons can be taught to detect specific input patterns. Neurons are activated when the specific input is provided to them which match the patterns that they were trained for. They consist of three groups or layers, of units: a layer of "**input**" units, "**hidden**" units and "**output**" units. Hidden Units are activated by the input units and the weights acting on them. The behavior of the output units is dependent on the hidden units and the weights connecting them. Neural Networks can have single-layer and multi-layer architectures with learning being supervised as well as unsupervised. Due to the inherent capacity of the neural network to learn in the presence of noise and handle non-linear data, they have found usage in bankruptcy prediction, speech recognition and fault detection.

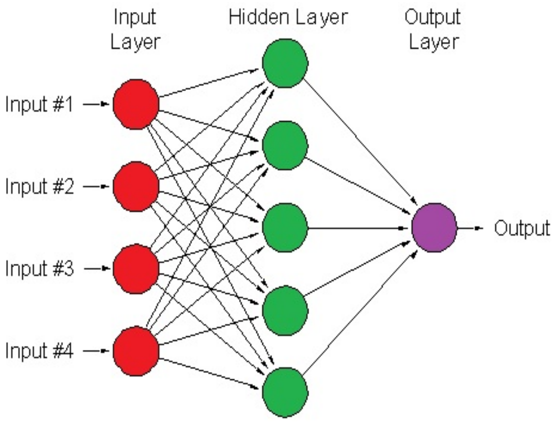

Figure 3. Neural Network

D. *Boosting Methods – Adaboost and Extreme Gradient Boosting.*

1) *Adaboost*

AdaBoost was the first boosting algorithm created for binary classification and has been used successfully using the Decision Trees Algorithm. Because these trees are so short and only contain one decision for classification, they are often called decision stumps. Each instance in the training dataset is weighted. All the observations have equal weight in the original training data set. Iteratively, each time classification is performed on the dataset, the observations that were incorrectly classified have their weights increased, while the observations that were correctly classified have their weights decreased. Predictions are made by calculating the weighted average of the weak classifiers. Adaboost has less parameters and no prior knowledge is required about the weak learners. However, it is prone to over fitting and may be vulnerable to uniform noise [11].

2) *Extreme Gradient Boosting*

Extreme gradient boosting brings together several weak "learners" into a single strong learner in an iterative fashion similar to gradient boosting. The goal is to "teach" a model to predict values in the form by minimizing the mean square error, averaged over some training set of actual values of the output variables. At each stage of the gradient boosting some imperfect model exists. The algorithm improves on the existing model iteratively each time adding an estimator in the process. Extreme Gradient Boosting also uses one hot encoding for the categorical variables which results in better interpretation in datasets which rely heavily on categorical data. The Extreme Gradient Boosting shows increased speeds and performance when used with parallelizable cores.

E. *Ensemble Methods – Random Forests*

Ensemble Methods are applied when a single classification model fails to deliver the required performance. Ensemble models are used to lower error rates and when the classification models are weak. Random forests are based on decision tree algorithms. Decision tree algorithms are capable of both classification and regression tasks on complex data sets. The decision tree can be for searching variable values pair in the training set and also create splits in such a way creating the best child subsets in the process using optimal splitting criteria, a process called tree growing. Conditions testing the attributes are used at every node to split the datasets. For classifying or making the predictions on the data we start at the root node and traverse down the branches till we reach the leaf nodes. The algorithm used to train a tree is called Classification and Regression Tree algorithm, also known as the CART. An Ensemble Learning thereby creates a strong learner using several decision trees as weak learners to achieve their goal.

V. DATASET DESCRIPTION

The data is sourced from the Poland’s Wroclaw Thoracic Surgery Centre and has been collected for patients who had severe lung resections for primary lung cancer. The dataset contains around 470 training examples of 16 features with each row indicating whether the patient survived as true or false. The data includes 3 continuous features and the remaining as categorical variables and is available in the UCI machine learning repository [12]. The samples have 85% negative cases and around 15% positive falling into the category of a perfectly imbalanced dataset.

VI. PERFORMANCE METRICS

For imbalanced data sets accuracy metric may not be suitable. Most common metrics used as Geometric Mean and F1 score which can help us evaluate the strength of the classifier for skewed data. In our work, we use True positive Rate (Sensitivity) and True Negative (Specificity) rate together with Geometric Mean as metrics to evaluate the performance of the algorithms. **True positive rate** also known as, **sensitivity** measures the proportion of positives that are correctly identified as the percentage of cases which are correctly identified as having the condition. **True Negative Rate**(TNR) also known as **specificity,** measures the proportion of negatives that are correctly identified as such as the percentage of healthy people who are correctly identified as not having the condition. The geometric mean G-mean is the product of the prediction accuracies for both classes, i.e. sensitivity and Specificity. As accuracy is a poor indicator of the performance of the algorithm on the positive samples, G-mean overcomes the problem since the value is dependent on both the positive and the negative samples in the dataset. However, Sensitivity still remains the focus of the study as we require better estimation of positive cases than the negative cases for healthcare data sets. Loss of accuracy in identifying negative cases may be tolerated for the above area.

$$TPR = TP\ /(TP + FN) \qquad (1)$$

$$TNR = TN\ /(TN + FP) \qquad (2)$$

$$Gmean = \sqrt[2]{\mathrm{TPR} * \mathrm{TNR}} \qquad (3)$$

TP is the true positives, TN is the True Negatives, FP is the False Positives and FN is the False Negatives.

VII. Results and discussion

Models were simulated in R using the R studio package. For Training phase, 70% of the data was allocated and the remaining 30% used for testing. Models selected for testing were the Naive Bayes, SVM, Neural Network, Adaboost, Extreme Gradient Boosting and Random Forest. The Performance of each of the predictive classifier is enlisted under two test conditions: without any pre-processing done to the datasets and with the SMOTE pre-processing applied to the dataset. Owing to the lesser number of positive samples, the models like the naïve bayes, Neural Network and Random Forest are unable to score on the sensitivity index under normal conditions (without SMOTE). The results of applying the various classifiers on the thoracic surgery dataset are provided below:

Method	**TPR**	**TNR**	**Gmean**
Naïve Bayes			
Without SMOTE	0.00	94.87	0.00
Using Smote	80.00	51.28	65.43
SVM			
Without SMOTE	10.00	97.43	31.21
Using Smote	**90.00**	48.72	**66.21**
Neural Network			
Without SMOTE	0.00	99.13	0.00
Using Smote	87.93	23.52	45.48
Random Forest			
Without SMOTE	0.00	98.71	0.00
Using Smote	40.00	60.25	49.09
Adaboost			
Without SMOTE	20.00	88.46	42.06
Using Smote	70.00	56.41	62.83
Extreme Gradient Boosting			
Without SMOTE	20.00	91.02	42.66
Using Smote	40.00	34.61	37.20

Table 1. Result under two Scenarios.

Hence the data imbalance issue has a significant impact on the performance of these models. Also tipping the imbalance ratio slightly towards the positive cases causes an improved performance of the models in identification of the positive cases. The model shows better performance than the models listed in [2]. Although using the classifiers on the bare dataset results in better TNR values, we get very poor TPR indicating that the classifiers are severely limited in identification of the much needed positive cases which is critical for the performance of the model especially in Healthcare Systems. All the models report a marked increase in the sensitivity metric when SMOTE is applied to the training data set and models such as Naïve Bayes, SVM and Neural Networks are able to identify the positive cases better than the other models. With the Geometric mean as evaluation criteria, SVM and Naïve bayes show significantly better performance amongst all the models.

VIII. Conclusion

In this paper, we have compared several Machine learning models using the imbalanced Thoracic dataset. We also evaluated and used metrics suited for imbalanced data on the classifiers under two test conditions: normal and with synthetic datasets. Our study reveals the effect of synthetic datasets on models and leads to positive cases being better identified leading to improved performance for the Healthcare dataset. The study can also be extended to analyze the effect of the SMOTE on other Healthcare datasets which are affected by the imbalance issue. Deep Learning methods may be utilized to further improve the Metrics as part of future work.

IX. References

[1] N.V. Chawla, K.W. Bowyer,L.O.Hall, W.P.Kegelmeyer, "Smote: syntheticminority over-sampling technique", Journal of Artificial Intelligence and Research, Vol.16. pp.321–357, 2002.

[2] Maciej Zieba, Jakub M. Tomczak, Marek Lubicz, Jerzy Swiatek, ”Boosted SVM for extracting rules from imbalanced data inapplication to prediction of the post-operative life expectancyin the lung cancer patients”,Applied Soft Computing,Vol 14,pp. 99–108,2013.

[3] M. Sultana, A. Haider, M. S. Uddin, “Prediction of Heart Disease using Machine Learning Algorithms:A survey",International Journal on recent and innovation trends in Computing and Communication , vol 5, issue 8, pp. 99-104, 2016.

[4] M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” Procedia Technol., vol. 10, pp. 85–94, 2013

[5] Hamouda S.K.M, Abo El-Ezz. H.R ,Wahed. ME,”Intelligent System for Predicting, Diagnosis and Treatment of Breast Cancer”, International Journal of Biomedical Data Mining,vol 6, Issue 2, 2017

[6] Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, van Berkel VH, Frieboes HB, “Prediction of lung cancer patient survival via supervised machine learning classification techniques”, Int JMedInform.,108:1-8. doi: 10.1016/j.ijmedinf.2017.09.013. Epub 2017.

[7] Savannah L. Bergquist,Gabriel A. Brooks, Nancy L. Keating , Mary Beth Landrum,Sherri Rose,”Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data”, Proceedings of Machine Learning for Healthcare, JMLR W&C Track Volume 68, 2017.

[8] Max Schubach, Matteo Re, Peter N. Robinson & Giorgio Valentini,” Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants”, Scientific Reportsvolume 7, Article number: 2959, doi:10.1038/s41598-017-03011-5 (2017)

[9] Nadir Mustafa, Jian Ping Li, Raheel A. Memon, Mohammed Z. Omer,"A Classification Model for Imbalanced Medical Data based on PCA and Farther Distance based Synthetic Minority Oversampling Technique", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 1, 2017

[10] Jiawei Han, Micheline Kamber, Jian Pei,"Data Mining Concepts and Techniques",3rd Edition, Morgan Kaufmann Publishers, pp.408-409,2012.

[11] Daniel.T.Larose, Chantal.D.Larose, "Data Mining and Predictive Analytics, 2nd Edition, Wiley Publications,pp.646-647, 2016-2

[12] M.Lichman,"UCI Machine Learning Repository", Irvine, CA: University of California, School of Information and Computer Science,2013.-3