Journal of Engineering Sciences
Vol 13, Issue 02, FEB /2022
ISSN NO:0377-9254

Imbalanced Datasets Classification Using Random Split Bagging Ensemble Technique

[1]S. Z. Parveen, [2]S. Nadiya, [3]C. Nancy

[1]Assistant professor Annamacharya Institute of Technology and Sciences Kadapa,

[2]Assistant Professor, KSRM College of Engineering

[3]Assistant Professor, Annamacharya Institute of Technology and Sciences-Kadapa

ABSTRACT

In Machine Learning, classification is considered as a supervised learning technique to predict class samples based on labeled data. Classification techniques have been applied on various domains such as intrusion detection, credit card fraud detection etc. However, classification techniques on all these domains have been applied on balanced datasets. Balanced datasets are those which contain equal proportion of majority and minority examples. However in real-time, obtaining balanced datasets is difficult because majority of the datasets tend to be imbalanced. Developing a model for classifying imbalanced datasets is a challenge particularly in medical domain. Accurate identification of disease affected patient within time is critical as any misclassification leads to severe consequences. However the imbalanced nature of most of the real time datasets presents a challenge for most of the conventional machine learning algorithms. For the past few years, researchers have developed models using machine learning algorithms which demonstrated unsatisfactory performance on classifying imbalanced datasets. As a remedy, few researchers have experimented with synthetic minority over sampling technique (SMOTE) and Cost Sensitive methods. Results indicated that these methods also have certain drawbacks such as over fitting and high mis-classificationrate.

To overcome these problems, ensemble techniques were proved to be robust in handling imbalanced datasets. In this study, we have considered existing boosting, bagging ensemble techniques and improved them in several aspects by proposing an algorithm named Random Split bagging such that imbalanced datasets can be handled effectively. This approach presented a novel tuple and attribute selection strategy. Finally we have chosen splitting criteria to generate class label.

The proposed random split bagging algorithm is compared with different machine learning techniques, data level methods, cost sensitive methods and ensemble techniques. Our proposed method traded-off good performance when compared with the state-of-art related works by achieving 98.8% accuracy. As the proposed model is experimented/validated with real-time datasets of imbalanced nature, this tool will be indeed immensely helpful to clinicians.

1. INTRODUCTION

Now-a-days people are affected by various kinds of diseases and diagnosing them has become a real challenge. Especially in the case of vector borne diseases like malaria and dengue, disease should be diagnosed immediately otherwise it may lead even to death of patient. Hence in diagnosing these type of diseases some level of expertise is required. Pathologists are the specialized professionals who will analyze the medical data samples and diagnose the disease. Pathologists need to assess data thoroughly before concluding the patient status. In several cases, it was observed that there are variations among pathologist's conclusions i.e. one pathologist conclude that patient has disease and other pathologist may conclude as patient not affected with disease. There was a conflict among decision on the same data. These variations are due to following reasons:

- Human vision/bias causing errors
- Lack of expertise
- Manual analysis complications.

Hence to eliminate the disadvantages of such manual diagnosis errors, there is a need for an automated system. According to Dr. Tedros Adhanom Ghebreyesus Director – General of World Health Organization (WHO) stated that use of computer based models in the medical domain are producing accurate results in analyzing the patient status. Automated systems are trained very well which are producing better results, helpful for complex data analysis and in-turn assisting doctors.

To analyze complex data with computer aided models, an expert artificial system is required. Machine learning is an application of Artificial Intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access and learn from data. But the primary goal is to allow the computers learn automatically without human assistance. The learning process basically starts with data collection or observations. For example, learning to diagnose the patients in medical domain, learning to drive an autonomous vehicle, learning to classify new astronomical structure's etc..

Methods or algorithms for learning by machines are often categorized into supervised learning, unsupervised learning, semi supervised learning and Reinforcement learning. This study focuses on *supervised learning* to classify disease affected patients. Supervised learning is also known as classification in which predictions can be done based on the labeled examples in the dataset. By nature, datasets can be typically categorized into two different type's namely *balanced datasets and imbalanced datasets*. Balanced Datasets are those whose class samples ratio is equally proportionate to each other where as in imbalanced datasets, one class sample ratio is dominates the other class sample ratio i.e. the class samples ratio are not proportionate to eachother.

A well-balanced dataset is suitable for the classifiers to perform classification efficiently. On the other hand, due to the unequal distribution of samples (imbalanced datasets) in a dataset, there exist certain issues such as

- Classifiers are biasing with majority classe xamples
- Cannot identify minority class example exactly

In most of the real-time cases, it is difficult to obtain balanced dataset. Specifically in medical domain, most of the datasets are of imbalanced nature. Learning on imbalanced datasets is a critical task because the majority class samples i.e., majority class sample data (unaffected patients) dominates the

minority class samples (affected patients data). Hence, during the learning phase, the classifiers selects only majority class samples for training because they are biased towards majority class samples and less trained with minority class samples. As a consequence, there is a chance of mis-classification of minority class samples. For instance, mis-classifying affected patient as un-affected Not only in medical domain there are so many other domains that are suffering with class imbalance problem. For example, intrusion detection, credit card fraud detections and detection of oil spills in satellite images etc.

Majority of the real-time datasets tend to be imbalanced. This study attempts to develop a model for classification of disease affected patients in case of medical datasets that are of imbalanced in nature.

2. LITERATUREREVIEW

In supervised method, classification is done using labelled data. Several automatic methods have been built to classify disease affected patients. However, all these methods proved to be robust in case of balanced datasets. Since majority of real- time datasets tend to be imbalanced, research has focused in the direction of disease prediction even in case of imbalanced data. In this section, we present research contributions made by various authors on explicitly handling imbalanced data and their learningmethods.

N. Poolsawad has experimented on LIFELAB dataset which was a clinically collected imbalanced dataset. Authors found that almost all *conventional data mining & machine learning algorithms* have a strong bias towards the majority class and are subject to error rate indicating poor recall. Jia Pengfei also reiterated that traditional classifiers are always showing poor performance in case of high density imbalanced datasets. To overcome biasing problem of conventional machine learning & data mining algorithms, researchers applied *data level methods* to handle skewed distributeddatasets.

Typically there are two data level approaches for handling imbalanced data sets namely *oversampling and under-sampling*. Over-sampling approach is defined as a method of balancing the minority class examples with majority class examples. In Under sampling method majority class examples are reduced in order to balance with minority class examples. Synthetic minority over sampling technique (*SMOTE*) is a popular method used for over sampling Rok Blagus et al. [9] investigatedSMOTE technique, conducted theoretical and empirical studies on sotiriou, Pittman and miller datasets and concluded that SMOTE is efficient for handling the low dimensional data when compared with high dimensional dataset. Also few researchers applied under sampling methods and compared performance of over sampling and under sampling techniques. V.Garcia et al. experimented with 17 highly imbalanced real datasets by applying both under sampling and over sampling techniques. Various classifiers namely Multi-Layer Perceptron-MLP, Support Vector Machine(SVM), Naïve Bayesian classifier (NBC), Decision tree- J48, Radial Basis Function(RBF), 1-Nearest Neighbor(1-NN), 7-Nearest Neighbor (7-NN), 13- Nearest Neighbor (13-NN) are applied by using weka toolkit with default parameter settings. Classifiers performance are evaluated with various performance metrics like True positive rate, True negative rate, index of balanced accuracy and Geometric mean. Finally authors concluded that over sampling methods are best for handling the imbalanced datasets when compared with performance of under samplingtechniques.

However, these over-sampling and under-sampling techniques have certain drawbacks. Loss of data is a major issue in case of under-sampling technique whereas over-sampling techniques are more inclined towards over fitting. It was clearly observed that the skewed distribution nature of imbalanced datasets is not changed either by over sampling or under sampling. Hence, research on solving the class imbalance problem taken into another direction i.e. concentrating on *algorithmic level* methods rather than on data level methods. Researchers proposed different algorithmic level methods such as *Cost sensitive classifiers and ensembles*

Salma Jamal et al. developed a Cost sensitive Naive Bayesian, Cost sensitive Random Forest and Meta Cost J48 algorithms on malaria disease dataset. Authors found that Cost sensitive Random Forest algorithm is the best predictive
model which has drastically improved hit rate. Gang-Song Xiao et al. developed a graph classification based on cost sensitivity (GCBC) and compared the computational complexity of GCBC with graph boost algorithm (gboost). Collected various graph datasets from carcinogenic potency database and proved that GCBC algorithm is efficient to handle graph based imbalanced datasets. However, cost sensitive classifiers suffer from high misclassification rate. Hence there is a need to improve accuracy levels of classifiers. To achieve this, *ensemble techniques* have been applied. Ensemble are a group of classifiers which work together *Voting* is a basic ensemble technique in which predictions are made based on grouping the decisions from all the classifiers with majority voted value. B. Romera-Paredes et al. has applied voting technique on accelerometric and gyroscopic data. Authors experimentally demonstrated that group of classifiers (ensemble) are performing better when compared to single classifiers. However, voting techniques failed to classify the imbalanced datasets as the base classifiers are traditional data mining and machine learning algorithms. Later research was driven in the direction of training to the classifiers on imbalanced datasets. For this, boosting and bagging ensemble techniques are found to be appropriate to train the classifiers

Shuo Wang et al. developed a model called Adaboost.NC for detection of software defects on various software datasets which are available from the public PROMISE repository. Taghi M. Khoshgoftaar et al. also found that bagging technique also performed well. Apart from these boosting and bagging techniques, few researchers used hybrid approach of classifiers i.e. combination ofclassifiers.

In adaboost, error of each classifier is carry forwarded to next classifier during its training phase which leads to more computations whereas random forest ensemble technique uses replacement strategy for selection of tuples in the process of classifiertraining. Hence, there is scope for improving adaboost and random forest classifiers in terms of tuple selection and attribute selection. Our study has considered this issue with an objective to improve the classification accuracy of imbalanceddatasets

3. METHODOLOGY

Handling imbalanced datasets is the challenge**.** To address this, under sampling and over sampling methods were designed. These methods suffer from data- loss and over-fitting issues. To resolve this, research has been shifted to another direction i.e. learning on imbalanced datasets at algorithmic level instead of data-level. Also cost sensitive methods were implemented. Experimental results indicated that mis-classification rate in algorithmic level methods is high.

Hence, neither cost- sensitive methods nor data-level methods performance are not superior in case of imbalanced datasets. In order to fix the issue, group of classifiers i.e. ensemble techniques have been proposed to handle imbalanced datasets. However, ensemble technique such as voting uses set of base classifiers. Since these base classifiers again are conventional ML algorithms, could not classify minority class samples effectively. Hence, we can conclude that data-level methods, cost-sensitive classifiers and voting ensemble techniques have not been found effectively working on imbalanced datasets. Therefore research was directed towards training classifiers in order to handle skewed datasets. Boosting and Bagging are the techniques that support this classifier training mechanism. In this study, we have considered existing boosting and bagging algorithms and improved them in several aspects by proposing an algorithm named *Random Split bagging* such that they can handle imbalanced datasets effectively. The methodology for proposed approach is discussed below.

Proposed approach:

Consider the proposed approach as shown in Figure 1, which improves the accuracy levels on classification of imbalanced datasets, especially accurate classification of minority class examples.

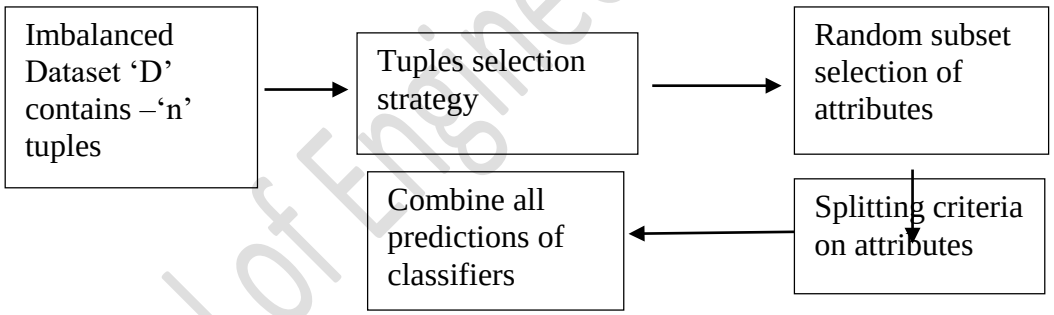

Fig.1. Proposed approach - Random Split Bagging ensemble technique

3.1 Imbalanced Datasets description:

In this research, three disease datasets related to malaria, dengue and jaundice were considered. Datasets are collected from medical wards of narasaraopet hospitals.

a) **Malaria disease Dataset :** The dataset consists of totally 165 patient records described with 13 different parameters of a patient namely Age, Hemoglobin, RBC, Hct, Mcv, Mch, Mchc, Platelets, WBC, Granuls, Lymphocytes, Monocytes as a set of input elements and status of Malaria – positive/negative as output element.

b) **Dengue disease Dataset:** The dataset consists of totally 68 patient records described with 2 different parameters of a patient namely platelets as an input element and status of Dengue – positive/negative as output element as stated below.

c) **Jaundice disease Dataset:** The dataset consists of totally 51 patient records described with 3 different

parameters of a patient namely Total bilirubin, Serum creatinine as a set of input elements and status of Jaundice – positive/negative as output element as statedbelow.

3.2 Tuples selectionstrategy:

In this step, initially we focused on the training the classifiers with skewed distribution of the dataset. Because of less number of minority class examples and more number of majority class examples, random forest ensemble technique selects the tuples with replacement during the tuple selection i.e. considering bootstrap sample to train the classifiers. This bootstrap sample selection process is susceptible to high bias. To resolve this issue, in this study we have modified tuple selection strategy. Tuples are selected randomly without replacement to train theclassifiers.

3.3 Random subset selection ofattributes:

After random selection of tuples without replacement from the dataset 'D', now the task is to improve the learning ability of the classifiers. In the case of imbalanced datasets, the training phase has to be done more effectively because the classifiers are tend to bias towards majority class examples or tuples when compared with minority class examples ortuples.

Hence, in order to improve classifier learning ability levels on imbalanced datasets, we focused on the selection of attributes in a tuple. In adaboost algorithm ensemble technique, classifier is trained without selection of subsets of the attributes.

During first iteration, randomly few tuples are selected *with all attributes* and used for training. This iteration will result a training error. This method was prone to error due to selection of all tuples. To minimize the error, our study proposes an attribute selection strategy. Initially tuples are selected randomly without replacement from a given dataset 'D'. Each tuple is defined with set of attributes and for each tuple, algorithm constructs subsets of the attributes randomly. Classifiers are trained with subset of the attributes that are selectedrandomly.

3.4 Splitting criteria on attributes:

Next step in the methodology of proposed ensemble technique is applying a splitting criteria which produces the accurate results on classification of examples especially on minority class examples. In this research, as splitting criteria on imbalanced datasets, "gini" index is applied on selected subsets of the attributes. The node with maximum splitting criteria is labelled first and the next maximum node value can places at the next level of tree and so on. The splitting criteria is used to classify the tuple or example belongs to a particular class.

3.5 Combine all predictions of classifiers:

Predictions are done based on combining the outputs from all the classifiers. Accurate classification of imbalanced dataset can be done by assigning the class label which is based on all the prediction values from the set of base classifiers.

4. Conclusion And Futurescope

4.1 Conclusion

Classification of imbalanced data with minority class samples (positive or affected patient) is a key area of research. So accurate prediction of a disease for a given patient and diagnosing within time is very important otherwise it may lead to aggravation of disease and finally leads to death of an individual. Also, it has

been observed that, many conventional classifiers are considering only majority class samples for classification ignoring minority class samples. This is due to imbalanced data that causes biasing problem. Conventional data mining and machine learning algorithms like Multi-Layer Perceptron, Linear Regression, Linear Discriminant Analysis, Radial Basis Function Network, Decision Tree – C4.5, Naïve Bayesian are biasing towards the majority class examples only and hence traditional data mining algorithms are not efficient to classify the imbalanced data accurately.

Since ML methods are not functioning accurately, data level methods namely SMOTE is applied in order to balance the class distribution of both majority (negative or un-affected) and minority (positive or affected) class examples. We applied SMOTE on datasets like malaria, dengue and jaundice disease datasets. However, we observed SMOTE (data level method) resulted in over-fitting. In order to overcome the drawbacks of data level methods, algorithmic level cost sensitive methods are applied. Cost sensitive classifiers namely *cost sensitive classifier – naïve Bayesian, cost sensitive classifier- random forest and Meta cost classifiers* are applied to classify the minority class examples. But cost sensitive classifiers have a drawback of high misclassificationrate.

To overcome with drawback of cost sensitive methods, researchers introduced ensemble techniques to improve the accuracy of classification on imbalanced datasets. In this study, three ensemble techniques namely voting, adaboost and random forest ensemble techniques are applied for classifying malaria, dengue and jaundice imbalanced disease datasets. For the classification of minority class examples voting method considers the classification results from all these group of classifiers and then assign the class label based on majority classification results. The drawback of voting ensemble technique is usage of traditional data mining and machine learning algorithms. Because of this, base classifiers are not trained properly due to the bias problem causing misclassification.

Later, boosting and bagging ensemble techniques i.e. Adaboost for boosting and random forest for bagging are considered to handle the imbalanced datasets. Adaboost ensemble technique in which it classifies the imbalanced datasets based on probability distribution and carrying the error produced by previous classifier to the next classifiers until error will be reduced. But, these sequential classification strategy is time taken process. Random forest ensemble technique is based on the selection of tuples with replacement i.e. depends on bootstrap sample for all the forest of classifier. Training to the ensemble of classifiers with bootstrap sample is computational expensive.

Hence, in order to improve the accuracy levels of classification on minority class examples over majority examples, a new methodology of ensemble technique is designed in this research that improves existing ensemble techniques. Proposed Random split bagging ensemble technique selects tuples without replacement and then to train the classifiers considering random subset selection of attributes from each tuple. Later, applied gini index splitting criteria for construction of classifier. Classification performance is evaluated by performance metrics like accuracy, precision, recall and F1-score. Also compared the performance of proposed random split bagging ensemble technique with other ensemble techniques, cost sensitive classifiers, data level methods – SMOTE, conventional data mining and machine learning algorithms. Proved that the proposed Random split bagging ensemble technique shows best performance on classification of minority class examples i.e.

classification of disease affected patients like malaria, dengue and jaundice diseases.

Further, the proposed random split bagging ensemble technique performance also tested for scalability on large sets. Experimentally we could found that our proposed approach is robust even with large datasets..

REFERENCES

1. *Thanh Quang Bui and Hai Minh Pham, "Web based GIS for spatial pattern detection: application to malaria incidence in Vietnam", Bui and Pham Springer plus, 5: 1014, pp. 1-14,2016.*

2. *Razali Tomaria, Wan Nurshazwani Wan Zakaria et al. "Computer Aided System for Red Blood Cell Classification in Blood Smear Image". International Conference on Robot PRIDE 2013-2014 - Medical and Rehabilitation Robotics and Instrumentation, ConfPRIDE 2013-2014. ELSEVIER Procedia Computer Science 42. pp. 206–213. 2014.*

3. *Jaree Thongkam, Gaundong Xu et al. "Toward breast cancer survivability prediction models through improving training space". ELSEVIER - Expert System with Applications 36. pp. 12200-12209.2009.*

4. *Yashasvi Purwar, Sirish L Shah et al. "Automated and unsupervised detection of malarial parasites in microscopic images". Malaria Journal 10:364. pp. 1-10.2011.*

5. *Francis Bbosa, Ronald Wesonga et al. "Clinical malaria diagnosis: rule-based classification statistical prototype". Springer Plus 5:939. pp.1-14.*

6. *N. Poolsawad, C. Kambhampati et al. "Balancing Class for Performance of Classification with a Clinical Dataset". Proceedings of the World Congress on Engineering. Vol I, pp.1-6.July2014.*

7. *Mikel Galar, Alberto Fern´andez et al. "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. pp. 1-22.2011.*

8. *Jia Pengfei, Zhang Chunkai et al. "A New Sampling Approach for Classification of Imbalanced Data sets with High Density". IEEE – Big Comp. PP.217-222.2014.*

9. *RokBlagusandLaraLusa."SMOTEforhigh-dimensionalclass-imbalanceddata". BMC Bioinformatics, 14:106. pp. 1-16. 2013.*

10. *V. García, J.S. Sánchez et al. "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance". ELSEVIER - Knowledge-Based Systems 25. pp. 13 - 21. 2012.*

11. *Yanmin Sun, Mohamed S. Kamel et al. "Cost-sensitive boosting for classification of imbalanced data". ELSEVIER - Pattern Recognition. pp. 3358-3378.Apr2007.*

12. *Salma Jamal, Vinita Periwal et al. "Predictive modeling of anti-malarial molecules inhibiting apicoplast formation". BMC Bioinformatics 14:55. pp. 1-8.2013.*

13. *Gang-Song Xiao and Xiao-Yun Chen. "Graph Classification with Imbalanced Data Sets". IEEE conference - The First Asian Conference on Pattern Recognition. pp. 57 - 61. Mar2012.*

14. *B.Romera Paredes, M.S.H.Aung et al. "A One-Vs-one Classifier Ensemble with Majority Voting for Activity Recognition". European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. pp. 443-448. April 2013.*

15. *Bartosz Krawczyk. "Learning from imbalanced data: open challenges and future directions". Prog Artif Intell. Springer. pp. 1-12. Apr2016.*

16. *Shuo Wang and Xin Yao. "Using Class Imbalance Learning for Software Defect Prediction". IEEE TRANSACTIONS ON RELIABILITY, VOL. 62, NO. 2, pp.434-443. JUNE 2013.*

17. *Taghi M. Khoshgoftaar, Jason Van Hulse et al. "Comparing Boosting and Bagging Techniques with Noisy and Imbalanced Data". IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans. VOL. 41, NO. 3. pp. 552 - 568. MAY20*

JES
Journal of Engineering Sciences

Vol 13, Issue 02, FEB /2022
ISSN NO:0377-9254