A Comprehensive Study on Ensemble-Based Imbalanced Data Classification Methods for Bankruptcy Data

K.UlagaPriya[12*], S.Pushpa[3]
Computer Science and Engineering
[1]Research Scholar, St.Peter's Institute of Higher Education and Research, Chennai
[2]Assistant Professor, Vels Institute of Science Technology and Advanced Studies, Chennai
[3]Computer Science and Engineering, St.Peter's Institute of Higher Education and Research, Chennai
[*]ulagapriya@gmail.com

Abstract— **In many real world classification problems the data is imbalanced where the distribution of classes is skewed. When the classification data are not approximately equivalent then the classification dataset is imbalanced. For example one class may be extremely low (minority class) and the other class may be extremely high (majority class). This imbalanced nature of data leads the prediction algorithm to be biased towards majority class. The poor representation of minority class affects the performance of the classification algorithm, which is evident through various assessment metrics. In this paper it is suggested to use ensemble techniques for imbalanced datasets, which focuses on binary class problems. The Ensemble bagging and boosting technique is applied on bankruptcy imbalanced data to improve the performance. Experimental study shows that better performance is achieved when SMOTEBOOST and SMOTEBAGGING are used with decision tree, which is a combination of SMOTE and Ensemble bagging and Ensemble boosting algorithm respectively and it outperforms other ensemble techniques.**

Keywords—Machine Learning, Imbalance data, Ensemble algorithm, classification

1. Introduction

The proportion of instances present in a dataset which is the class distribution plays a important role in classification problem. The skewed distribution of one class makes the data imbalanced since one of the classes has higher number of instances and the other class has lower no. of instances. For example in bankruptcy dataset the company who are bankrupt (minority) and the company did not bankrupt (majority) are the two classes in the dataset which is natural that the company who are bankrupt will be always low in the real world scenario. In this binary class dataset the minority class is low and hence when we apply a classifier, the performance gets biased towards majority class. In Literature several imbalanced classification problems are handled customer churn prediction [1], Autism Imbalanced data [2], diagnosing cancer[3], imbalanced health care data[4] has been analyzed in this context

One such Imbalanced binary class problem is bankruptcy prediction where the algorithm predicts the company is bankrupted or not. In this paper company bankruptcy dataset is taken from kaggle (Kaggle Website) where the data is imbalanced. Bankruptcy is a state of insolvency wherein the company or the person is not able to repay the creditors the debt amount. Hence it is necessary for any company to review its financial status and do prediction. Bankruptcy prediction is important not only to the society as such, but it is also important to the stakeholders. The purpose of the bankruptcy prediction of a company is twofold: a) Assess the financial status of the company and b) to focus on the future perspectives in the context of long-term operation. Bankruptcy prediction is trivial in the Finance Industry and it remains one of the hottest topics. The purpose of predicting financial distress is done through various economic parameters which enable the stakeholders to foresee the risk of Bankruptcy. Many approaches have been followed to predict bankruptcy. In this paper we approach ensemble bagging and boosting approach. Generally, methods which provide the solution to imbalance problem are grouped into sampling methods, active learning methods and cost sensitive methods.

Sampling methods: In this method class distribution is done by using various sampling Techniques like Random Over Sampling (ROS), Random Under Sampling (RUS), ADaptive SYNthetic sampling (ADASYN), Condensed Nearest Neighbor (CNN), Synthetic Minority Oversampling Technique(SMOTE), Edited Nearest Neighbor(ENN) and Neighborhood Cleaning Rule(NCL).This sampling method is applied on the training set and it is ensured that both the majority and minority class is balanced with equal distribution of instances. The drawback of this specific method is over fitting and it may cause the removal of necessary data.

Cost-sensitive methods: This method combines the approach of algorithm specific and data specific modifications for managing imbalanced data. When there is misclassification cost matrix is used. AdaC1, AdaC2, and AdaC3 were the methods proposed for cost sensitive approaches.

Active learning methods: Active learning efficiently chooses the instances from training data, so that computational cost is reduced when large Imbalanced datasets is involved. To handle the imbalanced training set, traditional active learning methods were used.

In this work, ensemble techniques are applied to handle the class imbalance issue. Apart from the above mentioned three methods, Ensemble Based techniques[5] can be applied to handle imbalance data. This technique is nothing but a combination of ensemble based techniques and any one of the methods specified above especially data level sampling method. This paper analyses the Bagging based Ensemble learning method and Boosting based Ensemble learning methods for bankruptcy Imbalanced dataset. This paper is ordered as follows. Section 2 analyses the Ensemble Techniques for Imbalanced data. Section 3 describes the experiments conducted and its details Section 4 presents the discussion based on the results of the experiments. Lastly, Section 5 illustrates the conclusion and future work.

2. Related Work

Previous work by authors are related to [6] various domains like software defect prediction, Credit card fraudulent prediction etc. [7] In Literature Ensemble algorithms like RUSBOOSTING, UNDERBAGGING, SMOTEBOOSTING, SMOTEBAGGING are used separately and hybrid approaches were also practiced with Imbalanced dataset. Different [8] approaches improve prediction based on the data distribution of the dataset and its nature. Literature proves that there is no best algorithm which gives best results for all the data sets. It purely depends upon the data that each algorithm works. Some algorithm gives best results for specific datasets.

3. Ensemble based Imbalanced Data Classification Methods

Ensemble based method is combined with different under and over sampling method which improves the performance of imbalanced dataset when compared to the traditional sampling techniques to handle with the imbalance data. This paper analyzes the [9]Bagging Based Ensemble Learning and Boosting Based Ensemble Learning to handle Imbalanced data.

Bagging Based Ensemble Learning

Bagging is nothing but Bootstrap aggregation is a powerful ensemble method bagging trains multiple classifiers in parallel. This makes differ set of bootstrapped replicas of original data set by randomly drawing different instances. Then apply the majority vote for prediction from different classifiers. This reduces the variance when it has high variance and approximately low bias, while retaining the low bias.

SMOTEBAGGING

SMOTEBAGGING is combination of SMOTE and BAGGING Algorithms. SMOTE (Synthetic Minority Oversampling) is an oversampling technique where synthetic data is generated for minority class to balance with the majority class. These synthetics data is generated based on K-Nearest Neighbor rule.

```
SMOTEBAGGING
Input : Dataset D with binary classes
Output : $D_{Train}$ Balanced Train data
For i 1 to N
        $\alpha = i/T$
    1.  Resample the majority class $D^{maj}$ with replacement at 100% as
        $D_r^{maj}$ = RandomSampleReplacement($D^{maj}$, $N^{maj}$ )

    2.  Resample the minority class with replacement based on $\beta$
        $D_r^{min}$ = RandomSampleReplacement($S^{min}$, $\beta$, $N^{min}$)

            Where $\beta = [ N^{maj} / N^{min} | ]\ \alpha$
    3.  Create new synthetic samples by means of SMOTE as
        $D_{new}^{min}$ = SMOTE($D_r^{min}$ , $\Theta$,K)

            Where $\theta$ =[ $N^{maj}$ / $N^{min}$| ] * [1- $\alpha$]
    4.  Construct $D_{train}$ = { $D_r^{maj} + D_{new}^{min}$ }
Endfor
    5.  Train a base-classifier ht → Y using $D_{train}$
```

UNDERBAGGING

Underbagging is combination BAGGING and Random under sampling and Algorithms. Random under sampling (RUS) is removal of instances from majority class which might also remove some important information.

```
UNDERBAGGING
Input : Dataset D with binary classes

Output : $D_b$ Balanced data

For i 1 to N
        $\alpha = i/T$
    1.  the minority class $D^{min}$with replacement at 100% as
        $D_r^{min}$ = RandomSampleReplacement($D^{min}$,, $N^{min}$ )

    2.  Resample the majority class with replacement based on $\beta$
        $D_r^{maj}$ =RandomSampleReplacement($S^{maj}$, $\beta$, $N^{maj}$)

            Where $\beta = [ N^{maj} / N^{min} | ]\ \alpha$
    3.  Undersample the data using Random unersapling as
        $D_{new}^{mAJ}$ = RUS($D_r^{maj}$ , $\Theta$,$N_{min}$)Where $\theta$ =[ $N^{maj}$ /
        $N^{min}$| ] * [1- $\alpha$]

    4.  Construct $D_{train}$ = { $D_r^{min} + D_{new}^{maj}$ }
Endfor

    5.  Train a base-classifier ht → Y using $D_{train}$
```

It keeps all minority data and randomly selects data from majority class which is equivalent to minority class.

SMOTEBOOSTING

SMOTEBOOSTING [10]is combination of SMOTE and AdaBoost.M2 Algorithms . Boosting provides equal weights to all misclassified samples. Boosting create ensembles. Ensembles, like bagging, boosting works powerful when the models are weak learner. Boosting algorithm samples from the data which are predominantly majority class subsequent sampling tends to biased towards majority class. Our aim is to reduce the bias by introducing SMOTE, so that while sampling in multiple rounds it enables to more no. of minority classes. This improves the overall accuracy by improving the True Positive.

```
SMOTEBOOST
Given: Set    $S = \{(X_1, Y_1), \dots, (X_m, Y_m)\}$ ,
  with labels  $Y_i$€ $Y = \{1,\dots C\}$

      where Cm, (Cm < C) indicates the
instances.

Let $B = \{(i,y) : i= 1\dots m \;\; Y \neq Y_i$
• Create the distribution $D_I$ over the instances, hence
      $D_1(i) = 1/m$.

• For t = 1, 2, 3, 4, … T
 1.Change the distribution $D_t$ by generating N synthetic data
 of minority instances
 $C_m$ by means of SMOTE algorithm
 2. Train a weak learner using distribution $D_t$
 3. Compute weak hypothesis
 4. Compute the pseudo-loss of hypothesis $h_t$:
 $\varepsilon_t =$      $\sum_{(I,Y\in B)} D_t(i,y)(1 - h_t(x_i, y_{i)}$

 5.          Set β and $w_t = (1/2)\cdot(1-h_t(x_i,y)+ht(x_i,y_i))$
  6. Update      $Dt : Dt + 1 (i,y) = ( Dt ( i, y ) / Zt )$
 where $Z_t$ is a normalization constant chosen such that $D_{t+1}$ is a
 distribution.
 7Output the final hypothesis
    •  $h_{fn=}$ arg max $\sum_{t=1}^{T} \log\left(\frac{1}{\beta}\right) h_t(x,y)$
```

RUSBOOST

RUSBOOST [11] is combination of Random under sampling and AdaBoost.M2 algorithms. The random under sampling reduces the majority class and makes the process faster. Whereas the boosting technique creates a composite classifier which ensures more accuracy than individual weak classifier. It assigns higher weights to misclassified samples .The RUSBOOST algorithm is similar to SMOTEBOOST algorithm except that instead of SMOTE, random under sampling is replaced. In SMOTE it identifies the K-nearest neighbors, whereas RUSBOOST it simply removes the instances randomly.

```
RUSBOOST
Given: Set $S = \{(X_1, Y_1), \dots, (X_m, Y_m)\}$ minority
class $y^r$ € Y }Y| = 2
    • Weak learner, Weak Learn
    • Number of iterations T

  1. Initialize
       a. $D_1(i) = 1/m$.

  2. For t = 1, 2, 3, 4, … T
  a) Create a temporary training set $S'_t$ with
     distribution $D'_t$ using random
     undersampling
  b) Call WekaLearn, Providing it with
     instances $S'_t$ and their weights $D'_t$
  c) Get back a
     hypothesis $ht: X \times Y \rightarrow [0,1]$
  d) Compute the pseudo-loss of hypothesis
     $h_t$:
     $\varepsilon_t$
     =
     $\sum_{(I,Y\in B)} D_t(i,y)(1 - h_t(x_i, y_{i)} + h_t(x_i, y))$

  e) Calculate the weight $\alpha_t = \varepsilon_t/(1-\varepsilon_t)$
  f) Update
     $D_{t+1}(i) = D_t(i)\alpha_t^{1/2(1+h_t(x_i y_i) - + h_t(x_i y \neq y))}$
  g) Normalize $D_{t+1}(i) ; Z_T = \sum( D_{t+1}(i))$
     $D_t : D_{t+1}(i) = ( D_{t+1}(i) / Z_t$

     Output the final hypothesis

  3. $h_{fn=}$ arg max $\sum_{t=1}^{T} \log\left(\frac{1}{\beta}\right) h_t(x,y)$
```

3. METHODOLOGY

This paper depicts the rebalancing technique using ensembles and then applying classification algorithm. Data is split into training set and test set. On the training data an ensemble technique is applied to generate a balanced training sample. The decision tree classifier is applied on the balanced training data and the performance of the classifier is evaluated against the test set. This process is repeated for different ensemble techniques and evaluated using performance metrics.

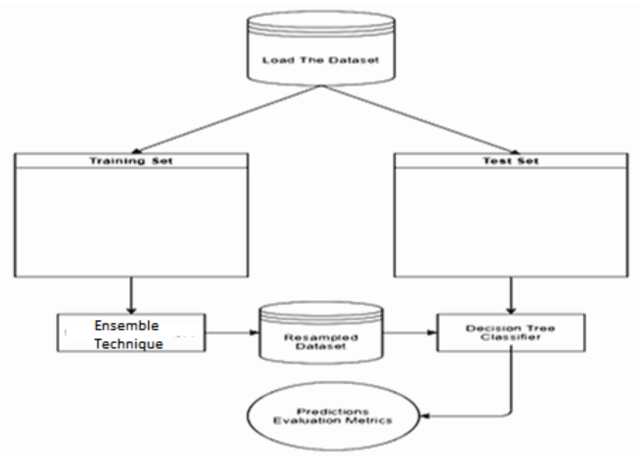


Figure 1 Flow Diagram

Dataset: The bankruptcy dataset is taken from kaggle where it has 65 attributes with 1000 sampes which is a binary class problem. This data set is imbalanced with imbalance ratio 0.02072063 where cases where bankruptcy prediction Yes and bankruptcy prediction No as minority and majority classes respectively.

Choice of classifier: The decision tree is taken as classifier to conduct the experimental study. The decision tree is appropriate for this experimental study for the following reasons:

- Decision tree effectively handles Imbalanced datasets.
- Decision tree squeeze all information especially when used in Ensemble algorithms
- Decision trees are ultimately weak learners. Though seems to be like a disadvantage, but Weak learners are efficient in the context of ensembles. Ensembles, like bagging, boosting works powerful when the models are weak learners

Performance metrics: The experiment study evaluates four performance metrics [12]than accuracy which are more suitable for Imbalanced data. Many researchers adhere to one metric but several metrics are available there is no best metric when considering Imbalanced data. The evaluation metric precision, recall and F1score are calculated based on True positive rate (TPR), false positive rate (FPR), and True Negative rate (TNR) and False Negative rate (FNR). AUCROC curve is another metric where a ROC curve is drawn FPR on x axis and with TPR on y axis.

Recall =True Positive /(True Positive+False Negative)

Precision=True Positive/(True Positive+False Positive)

F1 Score= Harmonic mean of precision and Recall

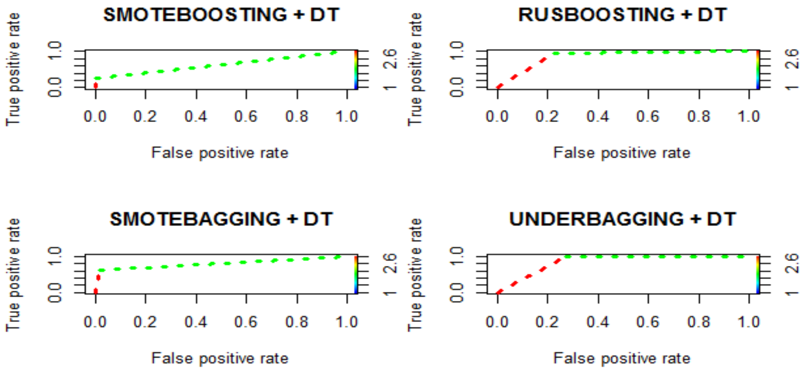
AUCROC indicates how well the probabilities from the positive classes are separated from the negative classes.

4. RESULTS AND DISCUSIION

The experiment was conducted in R with EBMC Package. The training set was run with Ensemble Algorithm+ Decision Tree classifier and a data model is created. With this data model test set prediction is done and performance metrics for bankruptcy data is evaluated. . The performance measures Recall, Precision, F1 Score AUCROC is calculated for each Ensemble algorithm – SMOTEBAGGING, UNDERBAGGING, SMOTEBOOSTING, RUSBOOSTING.

High recall, High precision: The classifier predicts most of the positive samples where Bankruptcy is yes, predicted also yes and Bankruptcy No, and then predicted also No. This enables both positive and negative cases to classify correctly. SMOTEBOOSTING AND SMOTEBAGGING techniques perform well with High Recall and High Precision.Overall, Recall depicts the negative samples where False Negative is high where the Bankruptcy were predicted correctly by Ensemble method using SMOTEBOOSTING AND SMOTEBAGGING

Algorithm	Recall	Accuracy	Precision	F-Measre
SMOTEBOOSTING	0.998	0.985	0.986922	0.99243
RUSBOOSTING	0.8059	0.8055	0.994329	0.890253
SMOTEBAGGING	0.9826	0.9705	0.987173	0.984881
UNDERBAGGING	0.7523	0.7555	0.997292	0.857643



CONCLUSION

Imbalance data is a challenging task in machine learning community. This bankruptcy data set is imbalanced with imbalance ration 0.02072063.To handle this imbalance Ensemble and sampling techniques SMOTEBOOSTING, RUSBOOSTING, SMOTEBAGGING and UNDERBAGING were applied on the training set with decision tree. The results give high recall for SMOTEBOOSTING and SMOTEBAGGING which is an important metric for Imbalanced data .It is also evident that SMOTE outperforms Random Under Sampling. Future research will be focused on large set of data tested with more number of Ensemble algorithms.

REFERENCES

[1] A. Amin *et al.*, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," *IEEE Access*, vol. 4, no. Ml, pp. 7940–7957, 2016, doi: 10.1109/ACCESS.2016.2619719.

[2] A. A. El-Sayed, M. A. M. Mahmood, N. A. Meguid, and H. A. Hefny, "Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE)," *Proc. 2015 IEEE World Conf. Complex Syst. WCCS 2015*, no. November, 2016, doi: 10.1109/ICoCS.2015.7483267.

[3] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Inform.*, vol. 90, 2019, doi: 10.1016/j.jbi.2018.12.003.

[4] I. Mohammadi, H. Wu, A. Turkcan, T. Toscos, and B. N. Doebbeling, "Data Analytics and Modeling for Appointment No-show in Community Health Centers," *J. Prim. Care Community Heal.*, vol. 9, 2018, doi: 10.1177/2150132718811692.

[5] P. Chujai, K. Chomboon, P. Teerarassamee, N. Kerdprasop, and K. Kerdprasop, "Ensemble Learning For Imbalanced Data Classification Problem," vol. 467, pp. 449–456, 2015, doi: 10.12792/iciae2015.079.

[6] S. Huda *et al.*, "An Ensemble Oversampling Model for Class Imbalance Problem in Software Defect Prediction," *IEEE Access*, vol. 6, no. c, pp. 24184–24195, 2018, doi: 10.1109/ACCESS.2018.2817572.

[7] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, 2012, doi: 10.1109/TSMCC.2011.2161285.

[8] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst. Man, Cybern. Part A Systems Humans*, vol. 40, no. 1, pp. 185–197, 2010, doi: 10.1109/TSMCA.2009.2029559.

[9] S. Ahmed, A. Mahbub, F. Rayhan, R. Jani, S. Shatabda, and D. M. Farid, "Hybrid Methods for Class Imbalance Learning Employing Bagging with Sampling Techniques," *2nd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2017*, pp. 1–5, 2018, doi: 10.1109/CSITSS.2017.8447799.

[10] N. V Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction," *Eur. Conf. Princ. Data Min. Knowl. Discov.*, vol. 2838, pp. 107–119, 2003, [Online]. Available: https://link.springer.com/content/pdf/10.1007%2F978-3-540-39804-2_12.pdf.

[11] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed," *Proc. - Int. Conf. Pattern Recognit.*, pp. 8–11, 2008, doi: 10.1109/icpr.2008.4761297.

[12] W. Feng, W. Huang, and J. Ren, "Class imbalance ensemble learning based on the margin theory," *Appl. Sci.*, vol. 8, no. 5, 2018, doi: 10.3390/app8050815.