Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 4- Issue 2, Paper 5, July 2021

A NOVEL FRAMEWORK TO IMPROVE THE EFFICIENCY OF THE CLASSIFIER ON HIGH IMBALANCED DATASETS

S. Babu,
Department of CSA,
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University,
babulingaa@gmail.com.

Abstract— *Class imbalance is one of the most challenging issues in the fields such as data mining and machine learning. The imbalance of data in the dataset arises when the data of each class is distributed unevenly. This occurs when the data associated with positive class is much smaller than the data associated with negative class. In this case, classifiers fail to identify the data of positive class. But, positive class is the key interest of the classifiers. Several solutions like sampling based models and hybrid models suggested to solve this issue. But all these are biased towards the undesired negative class, taking long training time, high storage and expensive computational issues. In addition, High dimensionalities in the dataset are the one another interesting challenges in machine learning. A dataset is referred to high dimensional when it has more dimensions or features. To reduce the dimension in datasets, feature selection technique is used which helps to improve the efficiency of the classifier in terms of accuracy and time. In this view, the proposed work defines model which improves the performance of the classifier in high dimensional imbalanced dataset. To prove the performance and efficiency of the proposed model, five imbalanced dataset comprising of both binary class and multi class are taken. Various experiments were done and the results were also compared. From the comparative results it is identified that the proposed model outperformed.*

Keywords — *Data Mining, Classification, Imbalanced Dataset, High Dimensionality, Feature Selection.*

I. INTRODUCTION

Data mining is the extraction of needful data from large databases and ignoring the rest. Data mining helps the people to make decisions on a situation as compared to statistical analysis [1, 7]. Data mining tools can pre-process the data and can work on unbalanced data easily. Data mining uses more direct approach and does meta-heuristics search on data. Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization [9]. The data mining techniques like clustering, classification, neural network, genetic algorithms help in finding the hidden and previously unknown information from the database. The purpose of data mining effort is normally either to create a descriptive model or predictive model. The purpose of predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable. If the target variable is one of a predefined number of discrete (class) labels, the data mining task is called classification [4]. The information or knowledge extracted so can be used for any of the following applications: Market analysis, Fraud detection, Customer Retention, Production control, Science exploration etc. The fundamental difference between data mining and statistical inference (such as sampling of datasets and proving of the null hypothesis) is that in data mining we mine huge amount of raw data and discover interesting patterns with which we can generate a hypothesis and answer certain questions [5,6].

Most of the real-world classification problems show some level of class imbalance, which is when each class does not make up an equal portion of the data-set. In the field of data mining and machine learning as most machine learning algorithms assume that data is equally distributed. Imbalanced learning occurs whenever some types of data distribution significantly dominate the instance space compared to other data distributions. In the case of imbalanced data, majority classes dominate over minority classes. Imbalance is usually referred in terms of ratio between the numbers of instances in the majority class to the number of instances in the minority class [12].

II. RELATED WORK

Nitesh V. Chawla (2002) proposed method SMOTE (Synthetic Minority Oversampling Technique). This algorithm combines oversampling and undersampling technique. To check the efficiency of this technique they use different kind of datasets and various classifiers like C4.5 decision tree, Naive Bayes, Ripper are used.

Yen, S. J., & Lee, Y. S. (2009) analysed various balancing techniques and presents four conclusions: Feature selection is a little better than sampling, When the dataset is largely imbalanced, undersampling is more useful, When the dataset is less imbalanced, they do not suggest pre-processing, In wrapper-based feature selection, complicated searching method may not get better results, for example, genetic searching performs worse than best-first searching.

Maisarah Zorkeflee, M., et.al, (2015) is proposed a technique to handle imbalance issue. It is the integration of Fuzzy Distance based Under Sampling (FDUS) and SMOTE. To evaluate the performance of the proposed method, measures like F-measure and G-mean are considered.

Jeatrakul, P., et.al, (2010) is proposed a method that combines both the SMOTE and the Complementary Neural Network (CMTNN) to solve imbalance issue. In this, CMTNN is functioned as an undersampling technique and SMOTE as an oversampling technique. To prove the performance, different datasets are considered. G-mean and AUC are used for analysis.

Maryam Zaffar & K.S. Savita (2018), presents the study of various feature selection algorithms and analysed their performance using two different datasets. This paper works with the feature selection technique with classification. The results indicated that there is significant performance difference of feature selection algorithms using the datasets with different numbers of features; shows 10 to 20 per cent difference in accuracy percentages. The performance of the filter feature selection techniques reduces as the number of feature increases. To predict the academic performance of the student, having a large number of feature sets, wrappers feature selection techniques can also be evaluated.

III. METHODOLOGY

In this section, the methodology was proposed in order to balance the dataset and to improve the performance of the classifier. In the proposed methodology, SMOTE (Synthetic Minority

Oversampling Technique) is used to balance the datasets. In SMOTE the minority class is over-sampled by creating “synthetic” examples rather than by over-sampling with replacement.

High dimensionality in datasets are one of the most important and interesting challenges in machine learning. A dataset is referred to high dimensional when it has more dimensions or features. To reduce the dimensions of the dataset a feature selection method is needed. In data mining various feature selection methods are available, from that two well-known methods are taken for analysis. A feature selection method called filter method and wrapper methods are compared and filter method is used to select the feature that has high information gain. The filter method calculates the gain value for each feature in the dataset. The least contributing features are removed. The most important features are selected by ranking the features. The classification is done with the datasets which are feature selected using filter method.

At first the methodology deals with the actual datasets which are imbalanced in nature and having multiple attributes, instance and features. In a multi-class dataset, majority class represents the class with highest number of instances and minority class refers to the class with the lowest instance levels. This causing the machine learning classifiers to be more biased towards majority classes. By using SMOTE, the dataset is balanced.

As a next step, Dimensionality is reduced. The dimensionality reduction is done through feature selection technique. In data mining various feature selection methods are available, from that two well-known methods are taken for analysis. A feature selection methods called filter method and wrapper methods are compared and filter method is used to select the feature that has high information gain because the filter method performs well on the high dimensional datasets in terms of accuracy and time. The attributes are selected using the Information Gain ratio with ranker method. The Information Gain value is calculated for each attributes in the dataset. After that attributes are ranked based on the Information Gain ratio. The attributes which have least score are rejected and the remaining attributes are taken for the classification. In feature selection methods the filtermethod is used for attribute selection. This methods is used for reduce the dimensions in the multidimensional datasets. The reduction of dimension reduces the time, by which improves the efficiency of the classifier.

Finally, the classification process is done. The datasets are compared with three classifiers to show the better performance in the classification process after the datasets are balanced. In the, proposed work, J48, Random Tree and Random Forest are the classifiers used to deal with the class imbalance problem. All the above three classifiers are result in very poor accuracy when the datasets are loaded as imbalanced. After the datasets are balanced using the oversampling technique SMOTE, required features are selected using feature selection, the classifiers are shown better performance in the classification process and the classifiers are efficient in terms of accuracy and time.

Algorithm for the proposed work:

```
Input: High imbalanced dataset.
Output: Balanced dataset.
  Step 1: Start.
  Step 2: SDS = SMOTE (IDS)
  Step 3: Datasets SDS with r Samples and k Features
  Step 4: For j = 1 to k
                FDS (j) = Filter (j)
          End
  Step 5: FDS1= high order sorting (FDS)
  Step 6: Selected Features = FDS1 (1 to k)
  Step 7: Classification using Selected Features
  Step 8: Stop
```

In the above proposed algorithm high dimensional imbalanced dataset taken as an input and returns low dimensional balanced dataset as an output. In the algorithm, dataset is denoted as IDS, samples as r, features as k and feature selection with filter method as FDS.

IV. RESULT AND DISCUSSION

In order to test the proposed method, five dataset from the UCI Machine Learning Repository and KEEL Repository were used. The Characteristics of the same are shown in Table 1. Ten cross fold method is used to split the data in to 90% of training set and 10% of testing set. To evaluate and to prove the efficiency of the proposed method, various types of test and comparison with existing methods has been carried out. They are,

1. Balancing the Imbalanced Datasets
2. Feature Selection
3. ROC Analysis of the proposed model

Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN : 2581 - 611X)

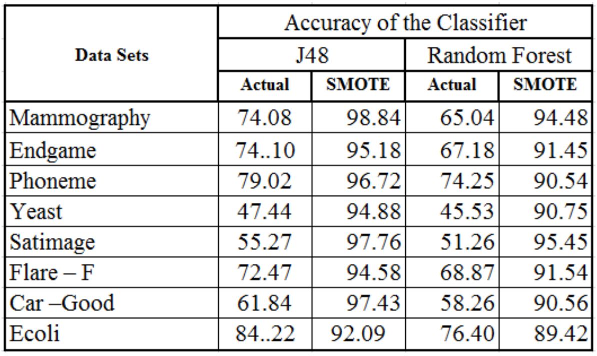
Table 1: Characteristics of the Datasets

Data Sets	No. of Instances	Minority Class %	Majority Class %
Mammography	1407	36.74	63.26
Tic-Tac-Toe Endgame	958	36.74	65.34
Phoneme	5404	29.35	70.65
Yeast	1485	28.96	71.04
Satimage	6435	9.73	90.27
Flare – F	1066	4.03	95.97
Car –Good	1728	3.99	96.01
Ecoli	353	7.36	92.64

4.1 Balancing the Imbalanced Datasets

Most of the datasets used for classification are imbalanced. The classifier will not work properly when the datasets are imbalanced. So, balancing the dataset is necessary to enable classifier to perform well. For balancing the datasets, different sampling techniques are exists. Out of which, SMOTE is one of the widely accepted oversampling technique to balance the dataset. In this view, in the proposed method SMOTE used. Using the WEKA tool, SMOTE method is applied on the above mentioned dataset to balance the same. After balancing, the classifiers are executed on the balanced dataset. The results are recorded and compared. The same is shown in Table 2.

Table 2: Balancing the datasets using SMOTE



Data Sets	Accuracy of the Classifier			
	J48		Random Forest	
	Actual	SMOTE	Actual	SMOTE
Mammography	74.08	98.84	65.04	94.48
Endgame	74..10	95.18	67.18	91.45
Phoneme	79.02	96.72	74.25	90.54
Yeast	47.44	94.88	45.53	90.75
Satimage	55.27	97.76	51.26	95.45
Flare – F	72.47	94.58	68.87	91.54
Car –Good	61.84	97.43	58.26	90.56
Ecoli	84..22	92.09	76.40	89.42

The results show that, the classifiers perform well on the dataset which was balanced by SMOTE. This in turn reveals that different classifiers focus on various techniques to improve their performance but no classifiers focused on the relative distribution of the each class in dataset.

4.2 Feature Selection

Feature selection is the process of selecting specific features from large no of features in a dataset. The filter method used for feature selection. Initially, the gain values for each feature in the dataset are calculated. .Based on the gain value, the least contributing feature is identified and removed. The most important features are selected by ranking the features. Feature selection will improve the performance of the classifier in terms of accuracy and improve the efficiency of the classifier in terms of time. The filter method of feature selection improves the accuracy of the classifier. The filter method of feature selection reduces the time taken by the classifier.

4.3 ROC Analysis of proposed model

The Receiver Operating Characteristic (ROC) Curves are the visualizing and summarizing technique of classifier performance. In ROC, point (0,0) states that classifier no way predicts positive instance, point (1,1) states that all instances are predicted as positive and point (0,1) is an ideal point that states that all positive instances are predicted as positive and no negative instances misclassified as positive. The Figure 1, shows the ROC curves of imbalanced and balanced E-coli dataset.

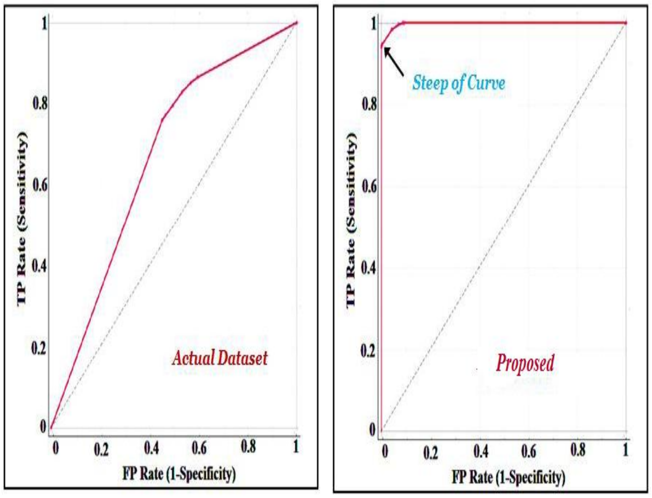


Fig. 1: ROC Comparison of E-coli Dataset

In addition, the steep of the curve also states that the number positive instances correctly classified are high and number of misclassified instances on negative class is very less when compared with actual dataset. This is the significant evidence which shows that; proposed model has greatest impact on the classifier performance in the context of imbalance presents in the relative distribution of each class.

V. CONCLUSION

In present day, class imbalance is a major problem in the research field. The dataset is said to be imbalanced when majority class has more instance than the minority class. The classifier works well on the balanced datasets. Imbalanced datasets can mislead a research work and also affect the efficiency of the classifier. All the existing classifiers are performing poorly on the imbalanced datasets. So it is very essential to balance the imbalanced datasets. In this research work the various methods available for balancing the datasets are analyzed and used the well-known technique called SMOTE to balance the datasets. Dimensionality Reduction is also another most important problem in machine learning. In order to reduce the dimension, feature selection technique is used. The feature selection technique will reduce the dimensions of the datasets by removing the non-impact features from the datasets. The filter method of feature selection is used to reduce the dimension of the datasets as it performs well than the wrapper method in terms of accuracy. Five imbalanced dataset with binary class and multi class are considered for analysis. Using SMOTE, the datasets are balanced and the same are evaluated in classifiers like J48, Random Tree and Random Forest with filter method of feature selection. From the analysis it is clearly found that the classifiers are able to achieve best performance and efficiency on the dataset which was balanced and dimension reduced by the proposed model. From the results of various experiments, it was concluded that balancing the dataset will lead to improve the performance of the classifier in terms of accuracy and also concluded that, reducing the dimension of the dataset will lead to improve the efficiency of the classifier in terms of time.

REFERENCES

[1] Aamer hanif, Noor Azhar, "Resolving Class Imbalance and Feature Selection in Customer Churn Dataset", IEEE International Conference on Frontiers of Information Technology, 2017.

[2] Arora, R, " Comparitive analysis of classification algorithm on different datasets using WEKA", Internationl Journal of Computer Applications, 2012.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall, SMOTE: Synthetic Minority Over sampling Technique, 16 321–357, (2002).

[4] Cios, K. J., Pedrycz, W., & Swiniarski, R.W. (1998). "Data mining and knowledge discovery", In Data Mining methods for knowledge discovery Springer US, 1-26.

[5] Deepika Tiwari, "Handling Class Imbalance Problem Using Feature Selection", International Journal of Advanced Research in Computer Science & Technology (IJARCST), Vol. 2, Issue 2, 2014.

[6] Maryam Zaffar, Manzoor Ahmed Hashmani, K.S. Savita & Syed Sajjad Hussain Rizvi, "A Study of Feature Selection Algorithms for Predicting Students Academic Performance", International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 9, No. 5, 2018

[7] Nadir Mustafa, Jian-Ping Li, "Medical Data Classification Scheme Based on Hybridized SMOTE Technique (HST) and Rough Set Technique (RST)", IEEE International Conference on Cloud Computing and Big Data Analysis, 2017.

[8] S. Babu, N.R. Ananthanarayanan, EMOTE: Enhanced Minority Oversampling TEchnique, Journal of Intelligent & Fuzzy Systems, 33 67–78 (2017)

[9] Yen, S. J., & Lee, Y. S. (2009), "Cluster-based under-sampling approaches for imbalanced data distributions". Expert Systems with Applications, 36(3), 5718-5727.

[10] Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm", In International Conference on Neural Information Processing, Springer Berlin Heidelberg, 152-159.

[11] Rahman, M. M., & Davis, D. N. (2013). "Addressing the class imbalance problem in medical datasets", International Journal of Machine Learning and Computing, 3(2), 224.

[12] Zorkeflee, M., Din, A. M., & Ku-Mahamud, K. R (2015). "Fuzzy and Smote Re-sampling Technique For Imbalanced Data Sets", Proceedings of the 5th International Conference on Computing and Informatics, ICOCI2015, Istanbul, Turkey, University Utara Malaysia.