Data Mining Techniques for Handling Imbalanced Datasets: A Review

P.Pavithra
Research Scholar
Department of CSA
SCSVMV University
Enathur, Kancheepuram
ppavithra.95@gmail.com

S.Babu, PhD
Assistant Professor
Department of CSA
SCSVMV University
Enathur, Kancheepuram
babulingaa@gmail.com

***Abstract* – In today's era of internet the amount of data generation is rapidly increasing. Some of the data related to medical, e-commerce, social networking, etc. are of great importance. But many of these datasets are imbalanced. A data set is called skewed or imbalanced when one of the classes highly dominates the others which results in poor efficiency of a classifier. To balance the imbalanced datasets we are going to analyse various methods available for balancing the datasets. About 52 research papers have been reviewed in this paper to identify the various data mining techniques to balance the imbalanced dataset.**

***Keywords* – Imbalanced dataset, Balancing data, Data mining techniques.**

1. Introduction

Most real-world classification problems showing 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. This causing the machine learning classifiers to be more biased towards majority classes. So it results in poor classification of minority classes. There are two types of imbalance learning problem.

Between class imbalances:

Imbalance that exists between the samples of two classes is usually known as between class imbalance.

Within class imbalances:

If the samples of the majority and minority are higher or lower than others then it is called within class imbalance.

This paper will identify and discuss the various methods for handling imbalanced data to improve the efficiency of the classifier.

2. Review of Literature

1. [Sotiris Kotsiantis, Dimitris Kanellopoulos and Panayiotis Pintelas] This paper describes various techniques for handling imbalance dataset problems. It discuss various techniques like undersampling, oversampling, threshold value, cost sensitive learning etc., It will give an overall understanding about the different methods available for pre-processing the data.

2. [Nitesh V. Chawla] In this paper they find a new 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.

3. [Puja Dwivedi, Udaya Kumar] In this paper they propose a new algorithm called Imbalanced Ensemble Feature Selection algorithm with sampled data using SMOTE. This will be designed to solve the low accuracy problem of minority classes, when on an imbalanced class dataset classifications are executed. They choose five datasets from UCI machine learning repository. They compare IEFS with other three classification algorithms. From the result of the experiments, they find that the predicting accuracy of minority class can be improved by the introduction of the attentions into diversity measure.

4. [H. Yin, K. Gai] This paper analyse various balancing techniques and presents four conclusions:

(1) Feature selection is a little better than sampling.

(2) When the dataset is largely imbalanced, undersampling is more useful.

(3) When the dataset is less imbalanced, they do not suggest pre-processing.

(4) In wrapper-based feature selection, complicated searching method may not get better results, for example, genetic searching performs worse than best-first searching

5. [S. Babu, N.R. Ananthanarayanan] In this work, they proposed new algorithm called Enhanced Minority Oversampling TEchnique (EMOTE) to solve the problem of imbalanced class distributions on the dataset. The experiments performed on nine different data sets using the machine learning algorithm C4.5 and other. The results of the classifiers is calculated in terms of effective measures like F-Measure, G-Mean and AUC and compared against various widely accepted methods. The analysis shows that the proposed EMOTE relatively generate a balanced dataset without any loss of information and without the inclusion of greater number of instances.

6**. [Reshma C. Bhagat, Sachin S. Patil]** In this work, the data pre-processing technique-SMOTE (Synthetic Minority Oversampling Technique) for multi-class imbalanced data is presented. Also, they have used RF (Random Forest) algorithm as a base, which is decision tree ensemble and known for its good performance. In today‘s scenario, big data is point of attraction because huge amount of data that are currently generating. Traditional data mining techniques are unable to survive with requirements urged by big data. They have tested quality of proposed system in terms of accuracy and F-measure. Experimental analysis carried out using various datasets of UCI repository. The results obtained shows that SMOTE+OVA algorithm gives good performance in the imbalanced data problem.

7. [T.Deepa, M.Punithavalli] This paper focused on extracting features from a high dimensional imbalanced dataset using EST (Evolutionary Sampling Technique) technique with a Genetic algorithm, and SVM (Support Vector Machine) to solve the misclassification and over fitting problem. In this proposed work two types of micro array dataset was taken which is naturally imbalanced and the results have been shown that the EST technique yields higher accuracy than random sampling technique.

8. [Nadir Mustafa, Jian-Ping Li] In this paper a new approach has been proposed for generating an accurate classification model, which aims to balance the minority instances in the training data. It can be done while applied the maximum distance based SMOTE, and also improve the accuracy of the testing data when applied the combined fuzzy roughest and SMOTE. From the qualitative and quantitative analysis different classifiers have been used based on the maximum distance SMOTE and it reveals that the new approach increases the performance of Area Under the Curve (AUC) metrics and accuracy metrics which used in a variety dataset. The present analysis shows that the Fuzzy Rough Set combined with MD SMOTE(Maximum Distance based SMOTE) technique is more effective than other combined approaches such as a Rough Set Theory (RST) and support vector machine (SVM).

9. [Sachin Subhash Patil, Shefali Pratap Sonavane] In this study, the enhanced data balancing techniques for two-class and multi-class imbalanced data have been presented using clustered based oversampling techniques. Various classifiers are used as base classifiers. The system quality testing benchmark may be indexed in terms of the parameters like accuracy, AOC area, G-Mean and F-measure. Experimental analysis is carried out using various data sets of the UCI/KEEL repository. The proposed three methods, namely MEMMOT (**ME**re **M**ean **M**inority **Over_Sampling** **T**echnique), MMMmOT (**M**inority **M**ajority **M**ix mean **O**ver_Sampling **T**echnique) and CMEOT (**C**lustering **M**inority **E**xamples **O**versampling **T**echnique) outperform the existing methods for selected data sets. The results indicate that proposed methods show an improved score of F-measure and ROC area, compared to the base techniques improving classification. The issues related to data set shift and changing oversampling rate needs to be further addressed in depth.

10. [Aamer hanif, Noor Azhar] In this research three methods were used to handle class imbalance problem. Three methods for feature selection have also been used. A random forest classification model was built to evaluate and compare the studied techniques. To study the impact of these methods, a real life dataset was used. Random oversampling gave the best result in this research to balance the datasets. However, all three feature selection techniques highlighted only the call related features as most important and performed equally well.

3. Current Approaches for Balancing Datasets

The literature survey suggests many algorithm and techniques that solve the problem of imbalance distribution of data. These approaches are mainly classified into four methods such as sampling, cost-sensitive learning, ensemble learning and feature selection.

3.1 Sampling Methods

We have basic sampling methods like oversampling, undersampling and advanced sampling methods like Tomek Link, Synthetic Minority Oversampling TEchnique (SMOTE), One-Sided Selection (OSS) Neighbourhood Cleaning Rule (NCL) Bootstrap-based Over-sampling (BootOS).

3.1.1 Undersampling: This_methods work by reducing the majority class samples. This reduction can be done randomly in which case it is called random undersampling or it can be done by using some statistical knowledge in which case it is called informed undersampling. Figure 1 shows the process of undersampling.

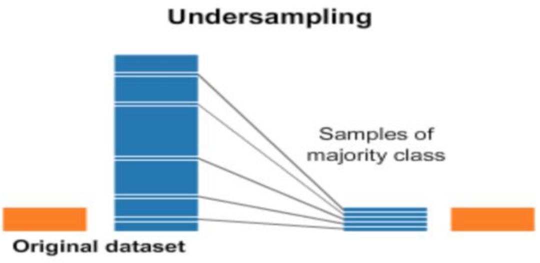

Figure 1: Undersampling

3.1.2 Oversampling: In this method, new samples are added to the minority class in order to balance the data set. Existing minority samples are replicated in order to increase the size of a minority class is called random oversampling. Figure 2 shows the process of oversampling.

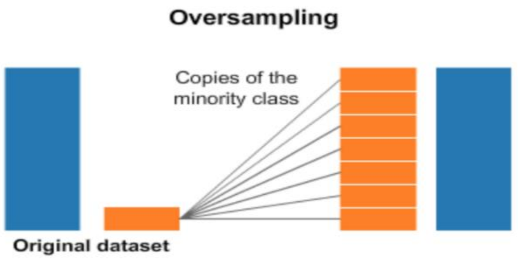

Figure 2: Oversampling

3.1.3 Synthetic Minority Oversampling TEchnique (SMOTE): In this method the minority class is over-sampled by creating “synthetic” examples rather than by over-sampling with replacement [2]. This is a statistical technique for increasing the number of cases in the dataset in a balanced way. The module works by generating new instances from existing minority cases that supplied as input. This implementation of SMOTE does not change the number of majority cases. . Figure 3 shows the process of SMOTE technique.

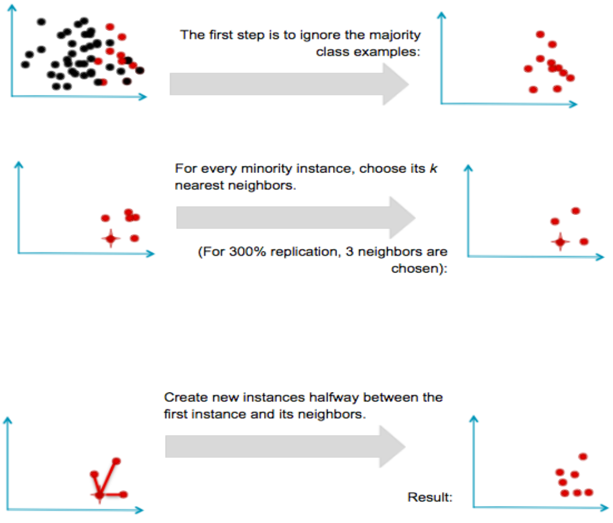
The first step is to ignore the majority class examples:
For every minority instance, choose its *k* nearest neighbors.
(For 300% replication, 3 neighbors are chosen):
Create new instances halfway between the first instance and its neighbors.
Result:

Figure 3: Synthetic Minority Oversampling Technique

3.1.4 One-Sided Selection (OSS): The One-Sided Selection is an under-sampling approach proposed by Kubat and Matwin [12], in which the redundant observations, noise, and boundaries are identified and eliminated from the majority class.

3.1.5 Using Tomek Link: Tomek link abbreviated TLink was proposed by Ivan Tomek in 1976 [13] as a method of enhancing the Nearest-Neighbor Rule. TLink can be used as a guide for under-sampling, or as a method of data cleaning.

3.1.6 Neighbourhood Cleaning Rule (NCL): It is an under-sampling approach that use the Wilson's Edited Nearest Neighbor Rule ENN to remove some observations from the majority class.

3.1.7 Bootstrap-based Over-sampling (BootOS): The BootOS is an over-sampling approach proposed by Zhu and Hovy[14]. Bootstrap set avoids exact duplication of observations in the minority class, and secondly, it can provide a smoothing of distribution on the training samples.

3.2 Ensemble Learning Methods

We have ensemble learning methods like bagging, boosting and random forest.

3.2.1 Bagging: It is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over fitting.

3.2.2 Boosting: Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning. It has various types of boosting like AdaBoost, Gradient Boosting and XGBoost.

3.2.3 Random Forest: Random forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. Balanced Random Forest (BRF) and Weighted Random Forest (WRF) are the two types of random forest.

3.3 Cost-Sensitive Learning

In many imbalance problems, not only the data distribution is skewed but also the misclassification error cost is uneven between the classes. The cost learning techniques take the misclassification cost in its account by assigning higher cost of misclassification to the positive class (minority class) and generate the model with lowest cost.

3.4 Feature Selection Methods

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection. It is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. It has various types like filter method, wrapper method and embedded method.

3.4.1 Filter Method: Filter methods are generally used as a pre-processing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. Figure 4 will describe the flow of filter method.

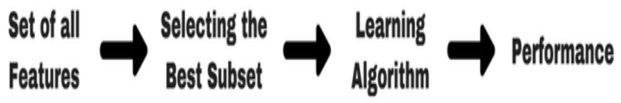

Figure 4: Process flow of filter method

3.4.2 Wrapper Method: It use a subset of features and train a model using them. Based on the inferences that drawn from the previous model, it decide to add or remove features from the subset. Forward feature selection, backward feature elimination, recursive feature elimination are some examples of wrapper method. Figure 5 shows the process flow of wrapper method.

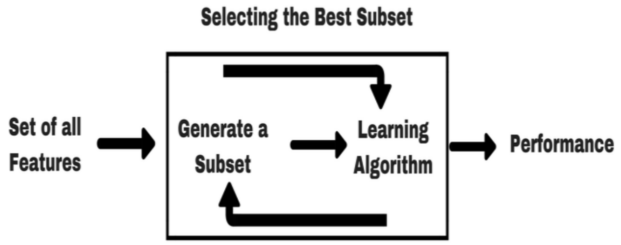

Figure 5: Process flow of Wrapper method

3.4.3 Embedded Method: This methods combine the qualities of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods. Figure 6 shows the process flow of embedded method.

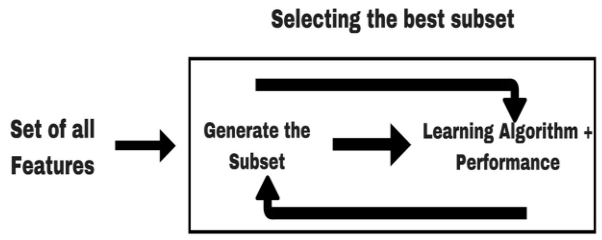

Figure 6: Process flow of Embedded method

4. Classification of Journals

A total of 52 journals have been selected and analysed in the above mentioned four dimensions. The objective of this analysis is to get the clear view on the following factors.

1) To investigate the level of research focus on the concept of data mining for balancing datasets.

2) To identified the techniques of data mining used for balancing the dataset to build an efficient classifier.

3) To study about the performance of the existing models.

4.1 Journals by Country of Authors

In this section the journals are year wise classified based on author of that journal who belongs to India. In table 1 out of the 52 papers reviewed only 10 papers have been published by Indian authors the rest were published by the authors of other nations. This will show our lack of contribution in balancing datasets.

Table 1: Year wise No of Journals based on authors country

Year of Publication	No. Of. Research Papers Based on Country	
	India	Others
2009	-	5
2010	-	2
2011	3	1
2012	-	4
2013	1	3
2014	-	7
2015	1	5
2016	2	4
2017	3	5
2018	-	6
Total	10	42

4.2 Journals by Techniques

In this section the journals are classified based on the technique used in those journals. In table 2 we compare 10 techniques with 52 journals to know the frequency of technique used for balancing the dataset. In that we can identify SMOTE is the most frequently used technique.

Table 2: No of Journals based on Technique

Techniques	No. of Research Papers
Undersampling	2
Oversampling	7
SMOTE	12
Hybrid sampling	7
Cost Sensitive Learning	2
Feature Selection	8
Ensemble Learning	7
Hybrid models	7

4.3 Journals by Techniques and Year of Publication

In this section journals are year wise classified based on the techniques used in those research paper. In table 3 we compare 10 techniques with 52 journals and the year of publication to know the frequency of technique used for balancing the imbalanced datasets. Though SMOTE is comes under the over sampling technique, we have separate column for that in the table because SMOTE is the most commonly used technique for balancing the imbalanced datasets. Hybrid sampling denotes the journals which used two or more sampling techniques. Similarly hybrid models denotes the journals which used two or more balancing techniques to balance the datasets.

Table 3: Journals based on Techniques and Year of publication

Year of Publication	No. of Research Papers							
	Under sampling	Over sampling	SMOTE	Hybrid Sampling	Cost Sensitive Learning	Feature Selection	Ensemble Learning	Hybrid models
2009	-	1	-	-	-	2	2	-
2010	-	-	1	-	1	-	-	-
2011	-	2	1	-	-	1	-	-
2012	-	-	2	1	-	-	-	1
2013	1	-	-	-	-	2	1	-
2014	-	2	2	-	-	-	2	1
2015	-	-	2	-	-	1	-	3
2016	1	-	2	-	-	-	2	1
2017	-	2	1	2	-	2	-	1
2018	-	-	1	4	1	-	-	-

5. Conclusion

In present day's class imbalance is a major problem in the research field. Imbalanced datasets can mislead a research work. In this paper we discuss the various methods available for balancing the data for efficient analysis. Many available techniques like sampling, cost sensitive learning, ensemble learning and feature selection are discussed in this paper for balancing the datasets. In this research work 52 research papers are reviewed and classified based on the author's country, year of publication and techniques used in those papers which helps to identify the frequency of the technique used for balancing the data. Of all the papers reviewed for this research SMOTE technique is most commonly used one and feature selection is the second most used technique. While comparing to other techniques the above two techniques give better results in balancing datasets. We also would suggest to pre-process the datasets before balancing, for better result. Out of the 52 papers reviewed only 10 papers have been published by Indian authors the rest were published by the authors of other nations. Still there is a lot more to contribute for this topic, so I would be working more on this topic of research.

6. Limitations of Research

In this work not all the research journal are considered for the review. But, only research journals which are reliable in all the perspective are considered. So this review has some limitations.

□ As first, 52 journals from the last one decade has been considered for the study. More journals can be selected for research.

□ As Second, the journals were searched based on the technique used for balancing the datasets of some real-world problems. More problems and techniques can be studied and analysed for further research.

References

[1] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, Handling imbalanced datasets: A review, GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006

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

[3] H. Yin and K. Gai. An empirical study on pre-processing high dimensional class-imbalanced data for classification. In 2015 IEEE 17th International Conference on High-Performance Computing and Communications; The IEEE International Symposium on Big Data Security on Cloud pages 1314–1319, New York, USA, 2015.

[4] Puja Dwivedi, Udaya Kumar, A Review on Classification Algorithm for Imbalanced-Class Datasets, International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 6, Issue 5, May 2017

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

[6] Reshma C. Bhagat, Sachin S. Patil, Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data using Random Forest, IEEE, 2015

[7] T.Deepa, M.Punithavalli, A New Sampling Technique and SVM classification for Feature selection in High-Dimensional Imbalanced dataset, IEEE, 2011

[8] 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

[9] Sachin Subhash Patil, Shefali Pratap Sonavane, Enriched Over_Sampling Techniques for Improving Classification of Imbalanced Big Data, IEEE Third International Conference on Big Data Computing Service and Applications, 2017

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

[11] Apurva Sonak, R.A.Patankar, A Survey on Methods to Handle Imbalance Dataset, IJCSMC, Vol. 4, Issue. 11, pg.338 – 343, November 2015.

[12] Kubat M, Matwin S, "Addressing the curse of imbalanced training sets: One-sided selection", In Douglas H. Fisher, editor, ICML, pages 179–186, 1997.

[13] Tomek Ivan, "An Experiment with the Edited Nearest-Neighbor Rule", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 6, No. 6, pp. 448-452, 1976.

[14] Zhu Jingbo, Hovy Eduard, "Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem"; Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 783–790, (2007).