A Hybrid Combined Under-Over Sampling Method for Class Imbalanced Datasets

Dr.M.Vijayakumar[1], E.Prabhakar[2]

[1]Professor, Department of CSE, Nandha College of Technology, Tamil Nadu, India, Email: tovijayakumar@gmail.com

[1]Assistant Professor, Department of CSE, Nandha College of Technology, Tamil Nadu, India, Email: prabhakarit10@gmail.com

Abstract - Imbalanced data sets can be balanced by using sampling methods like under sampling and over sampling. But still, both the sampling methods have some drawbacks. This review proposes the under-over sampling approach could overcome drawbacks of the under sampling and the over sampling methods. At first, Under-over sampling method combines the basic ideas of the under sampling and over sampling methods. Sampling method introduced is majorly uses the improvement of predicting accuracy of minority class(es) for two and multi class imbalanced datasets. Results show that the new sampling method improves the prediction accuracy of minority class(es).

Keywords - Classification, Sampling, Under Sampling, Over Sampling, AdaBoost, Imbalanced datasets, Balanced datasets, Two class imbalance, Multi class imbalance.

I. INTRODUCTION

The process of data mining involves analyzing the entire datasets, understanding the properties of datasets and then makes classification, association or clustering. Classification is one of the hot research areas in the field of data mining. Due to the prediction ability of classification, it has been used in many of the real applications. Classification builds a classifier with set of rules by analyzing the entire datasets and understanding the properties of datasets and then this set of rules are applied to the new datasets [1]. For the two class datasets, the overall accuracy is total number of instances correctly classified divided by total number of instances. A good classification algorithm should have high prediction accuracy and low prediction error for all classes. Because sometimes overall prediction accuracy will be very high or more than 95% but prediction accuracy of one class will be very low or sometimes 0%. In these cases, based on overall prediction accuracy we cannot assume that the particular classification algorithm is good one.

Consider two class dataset with the ratio of two classes are 10:90. Assume 2 algorithms used. The first algorithm gives out the overall prediction accuracy value about 90% having accuracy of all class is 0% and 100% correspondingly. The second algorithm brings out the accuracy as 78% which is an entire prediction with accuracy for every class namely 60% and 80% correspondingly. By compare the entire accuracy of prediction for both algorithms, the first one obtains high accuracy than second one. Certainly, first classification algorithm could misclassify all of examples in the class1. Finally, by means of this conclusion, first classification approach is not assumed as good classification scheme. These types of datasets are called as imbalanced datasets and many of the real applications such as medical research[2] and fraud detection having imbalanced datasets. Minority class instances are misclassified into majority class instances in imbalanced datasets. It happens due to number of instances in majority classes are very high compared to number of records in minority class. So, classifier is trained heavily based on the majority class instances and classifier ignores minority class instances while training. Because of this problem, classifier predicts the majority class instances correctly and minority class instances incorrectly. If imbalanced datasets are converted into balanced datasets, then classifier will be trained for all classes equally and it will predict all instances correctly and misclassification in minority class instances will be reduced [18]. The traditional data mining algorithms generally assumes that the given dataset is balanced dataset. Based on this assuming it makes the classification. Thus traditional data mining algorithms are not suitable for imbalanced datasets [17].The rest of the paper is organized as follows. Section II discusses the related works. Section III overviews the proposed system methodology and gives detailed experimental results of the proposedscheme. Concluding remarks and future works are covered in Section IV.

II. RELATED WORKS

Imbalanced datasets can be balanced by two approaches. First one is data level approaches and second one is algorithm level approaches [12].

A. Data Level Approaches

The approaches at data level try to provide the solution to imbalanced datasets by converting imbalanced datasets into balanced datasets. These approaches create balanced datasets at data level from imbalanced datasets.

Under Sampling shortly (*US)* minimize the amount of instances in the majority class by means of random removing of instances or making clusters and removing the instances which is depend on clustering[6].

Random Under Sampling simply (*RUS)* reduces the instances availability in majority class with the process of randomly remove of instances from the majority class [6].

Cluster Based Under Sampling(SBC) first clusters all the instances into some clusters. Then from each cluster, it selects suitable number of instances from majority class and suitable number of instances from minority class by taking the ratio of number of majority class instances into number of minority class instances in the cluster [19].

One Sided Selection (OSS) balances the dataset by eliminating the noise and redundant instances from the majority class [6].

Condensed Nearest Neighbor Rule (NCL) eliminates the instances from the majority class that are distant from the decision border, because these instances might be considered as less relevant for learning [6].Neighborhood Cleaning Rule also said to be as NCL which use the Wilson's Edited Nearest Neighbor Rule called ENN that removes the majority class instances. ENN would remove any of instance where class labels could differ the class in at least 2 of their 3 nearest neighbors [6].

Tomek Links (TL) removes noisy and borderline instances from majority class based on tomek link formation [6].

Advantages There is no chance of over fitting and Under Sampling is less time consuming method.
*Disadvantage*Under Sampling will lead to loss of information.

Over Sampling (OS) increases the number of instances in minority class by replicating or synthetically generating instances to reduce the level of class imbalance [6].

Random Over Sampling (ROS) adds instances to minority class by randomly replicating instances in minority class [6].

Synthetic Minority Over sampling TEchnique (SMOTE) adds instances to minority class by synthetically generating instances from minority class. It avoids over fitting but lead to over generalization because SMOTE generates synthetic minority class instances without considering majority class instances [3,8].

Borderline SMOTE shortly BSM selects the minority class instances in border of minority decision region in feature space and the SMOTE algorithm could be applied for over sample the instances instead of over sampling the all [11].

Advantage There will be no loss of information while Over Sampling methods are used.

Disadvantages The Over Sampling approach would leads to the over fitting and that would take much time due to adding instances.

B. Algorithm Level Approaches

The approaches at algorithm level try to provide the solution by adapting existing classification algorithms to increase the training regards to minority classes[4].

AdaBoost improves weights of the instance that are misclassified and would reduce the weights of those instances which are classified correctly [10]. It does not consider the imbalance but it slightly more concentrate on minority class instances compared to other classification algorithms. In [15] authors introduced New Adaboost Approach for Sentiment Analysis of US Airline Twitter Data.

Balanced Random Forest (BRF) is the improvement of random forest algorithm, specially designed for imbalanced datasets. All iteration in random forest, draw a bootstrap sample from the minority class. Randomly draw the same number of cases, with replacement from the majority class. Then finally aggregate the prediction to make final prediction [7].

C. Combination of data and algorithm level approaches

SMOTE Boost provides integration of SMOTE into a standard boosting algorithm. First over sampling is performed using SMOTE and then standard AdaBoost algorithm is applied [9].

RUS Boost applies the RUS first, is an approach which randomly removes the examples from majority class to the balance class distribution and it applies on standard AdaBoost [13].

Exactly Balanced Bagging (EBB) combines the entire minority class with random subset of the majority class and then it applies the bagging algorithm [21].

Over Bagging procedure uses over sampling of minority class instances and then bagging concept is applied to the classifier by assigning equal weights [14].

Under Bagging procedure uses under-sampling of majority class instead of over-sampling minority class. The under-sampling steps are applied for majority class and now bagging concept is applied on classifiers by means of assign the equal weights [5].

The data collected for public opinion mining of government scheme advertisement, contains class imbalance. Authors in [16], [21] handled class imbalance with new novel approach.

III. PROPOSED METHOD & EXPERIMENTAL DESIGN

Two various under over sampling mechanisms are proposed to handle imbalanced datasets in 2 and the multi class environment.

A. Two class imbalance datasets

In the two class imbalance environment, there will be only two classes. One class will have very large number of instances while another class will have very small number of instances. So defining majority class and minority class is easy in two class imbalance environment, unless both the classes have equal number of instances in each class. Only very rare number of applications will have equal number of instances in each class. The class that possesses the large amount of instances has assumed as majority class and that classes having little number of the instances has considered as one of minority class.

Structure of under-over sampling method for two class imbalance environment

1. Calculate the overall number of instances and then find the average number of instances.
2. Consider that average number of instances as a sampling value.
3. The class having high amount of examples than the sampling value is assumed as majority class and the instances in class which are under sample by means of random under sampling.
4. Under sampling is performed until it reached the sampling value.
5. The class could comprise fewer instances than the sampling value is considered as one of minority class and those instances in class are to be over sampled by random over sampling.
6. Over sampling is performed up to it reach the sampling value.

Algorithm

- Let N-Number of Instances, C-Number of Classes
  Majority Class=MC, Minority Class=MiC
- Calculate Average Number of Instances A=N/C
- For Majority Class Instances,
  Apply RUS
- Apply until MC=A
- For Minority Class Instances,
  Apply SMOTE
  smote= floor((2 * size_N - size_input)/size_A)
- Apply until MiC=A

Two class imbalance environment

Consider the two class imbalanced dataset with 10,000 instances. One class (take it as class a) having 7500 instances. Other class (take it as class b) having 2,500 instances. Here, defining majority class and minority class is very easy. Class a has very large number of records and class b has very small number of records. So, a class has been considered as the majority class and class b as a minority class.

Averagely, 5000 records are there and average of that is assumed as sampling value. Instances in class a are applied for under sampled that uses random under sampling until number of instances in the majority class which reaches the

sampling value of those 5000. Likewise, instances of class b would over sampled with the help of random over sampling up to instances in minority class which would reach sampling value over 5000.

In the imbalance scenario, class a has 7500 instances and in balance scenario class a has 5000 instances (7500 (original)-2500(under sampled) =5000). Similarly in the imbalance scenario, class b has 2500 instances and in balance scenario class b has 5000 instances (2500 (original) +2500(over sampled) =5000).

If the imbalance datasets are not balanced, then many of the minority class instances will be predicted as a majority class instances. Because of classifier might be trained heavily based on majority class instances and rules will be created based on training from majority class instances 0f 7500 and minority class instances of 2500.

If imbalances datasets are get balanced, both majority class instances and minority class instances would train equally and the rules would create depending on training by means of majority class and minority class instances of that 5000. Finally, the prediction is not found as favor for majority class instances or it may favor for both.

B. Multi class imbalance datasets

In multi class imbalance, it is identified that those would have more than two classes. One class or two classes will have very large number of instances while another class (es) will have very small number of instances. So defining majority class and minority class is difficult in multi class imbalance environment.

Considering three class imbalance environment, it possess three classes such as a, b, c. Among these the class a is majority class which is respective to class b but similar class a will minority class with respect to both class c and b. Here class a could considered as majority with respect for class b but minority class for class c.so defining class range for class a will be difficult in this scenario. There is a need for some method to define majority class and minority class.

Structure of under-over sampling method for multi class imbalance environment

1. Calculate the overall instances and average number of instances in the multi class datasets. Then consider that average number of records as a sampling value.
2. The class which has higher capacity of instances than the sampling value had considered as majority class and the one possess less instances than sampling value is assumed to be as minority class.
3. The instances in majority class can be under sampled using random under sampling. In case, larger than the one majority class, then all are made to under sampled by means of using random under sampling.
4. Under sampling is performed on majority class (es) untill majority class instances reach the sampling value.
5. With the help of random over sampling, the instance in specific minority class is over sampled. If they found as higher than particularly one minority class, then all will be over sample by means of random over sampling.

Algorithm

- Let N-Number of Instances, C-Number of Classes
  Majority Class=MC1,MC2,…., Minority Class=MiC1,MiC2,….
- Calculate Average Number of Instances A=N/C
- For Majority Class Instances,
  Apply RUS
- Apply until MC1=A, MC2=A,….
- For Minority Class Instances,
  Apply SMOTE
  smote= floor((2 * size_N - size_input)/size_A)
- Apply until MiC1=A, MiC2=A,…..

Three class imbalance environment:

Consider the three class imbalanced dataset with 10,000 instances. One class (take it as class a) having 1500 instances. Other class (take it as class b) having 3500 instances and another class (take it as class c) having 5000 instances. Here, defining majority class and minority class is very difficult. Because here class b is considered as majority class with respect to class a. But it is considered as a minority class with respect to class c.

Average records of 3333 are used and their average is taken for as a sampling value. Here class a is minority class, since it possess very little amount of instances namely 1500 than the sampling value. In similar way, class b is considered as majority class, thus it has the large instances such as 3500 than sampling value and class c is the majority class, due to the reason that it possess the high number of instances called 5000 than specific sampling value. In this situation, two

majority and one minority classes are found and used. The majority class that contains much instances than other majority one is consider as a majority class 1 and the another majority class as majority class2.

In case, if two minority classes are there, then the minority class with very small number of the instances than another is consider to minority class 1 and other as a minority class2. Any of those two of classes with same amount of instances in every class, then we assume by their own way, which could suitable. Simply some rare application could have this kind of situation as two or more classes contains of same instances in all the class.

Instances of class a are gets over sampled by means of random over sampling until the instances in minority class reaches sampling value namely 3333. Similarly instances in class b are get to be under sampled with the use of random under sampling up to number of the instances in majority class that reach the specified sampling value of about 3333 and all instances in class c are get under sampled using the random under sampling up to amount of instances in specific majority class which reaches that sampling value such as 3333.

In the imbalance scenario, class a has 1500 instances and in balance scenario class a has 3333 instances (1500 (original)+1833(over sampled) =3333). Similarly in the imbalance scenario, class b has 3500 instances and in balance scenario class b has 3333 instances (3500 (original) -167(under sampled) =3333) and class c has 5000 instances and in balance scenario class c has 3333 instances (5000 (original) -1667(under sampled) =3333).

If the imbalance datasets are not balanced, then many of the minority class instances will be predicted as a majority class instances. Because of classifier might be trained heavily based on majority class (es) instances and rules will be created based on training from majority class instances of 5000&3500 and minority class instances of 1500.

If the imbalance datasets are balanced, then both the majority class instances and minority class instances will be trained equally and rules will be created based on training majority class instances of 3333 & 3333 and minority class instances of 3333. So prediction will not favor the majority class instances, it will favor the all the classes.

EXPERIMENTAL DESIGN

A. Dataset preparation

Three two class datasets abalone, census &vote with two multi class datasets nsl& balance are taken from weka, keel – data mining tools & UCI data repository. For the multi class dataset, three class dataset is considered. All the datasets consist of class imbalance problems in different ratios.

TABLE 1: TWO CLASS DATASETS CHARACTERISTICS

Dataset	Attributes	Overall instances	Instances in class 1	Instances in class 2	Ratio of class imbalance (class 1 : class 2)
Abalone	4	731	689	42	94.25 : 5.75
Census	7	692	657	35	94.94 : 5.06
Vote	5	335	267	68	79.70 : 20.30

TABLE 2: MULTI CLASS DATASETS CHARACTERISTICS

Dataset	Attributes	Overall instances	Instances in class 1	Instances in class 2	Instances in class 2	Ratio of class imbalance (class 1 : class 2 : class 3)
NSL	6	8896	4342	2152	2402	48.81 : 24.19 : 27
Balance	5	625	49	288	288	7.84 : 46.08 : 46.08

B. Performance Metrics

Overall prediction accuracy has been used to evaluating the performance of specific classifier algorithm on the imbalanced datasets, thus overall prediction accuracy could be high in sometimes, but will perform badly in minority class and performs good in the majority class. Thus overall prediction accuracy is not accepted performance metric of the imbalanced datasets.

If datasets are high imbalanced, even the classifier is classifies all of majority class instances by correctly and misclassifies all minority class instances, then the overall accuracy of classifier is still found as large due to since it posses high majority class instances than the minority class instances. Under this scenario, overall prediction accuracy will not reflect the reliable prediction for minority class. So, separate accuracy for minority class and majority class are needed.

Overall prediction accuracy is number of correctly classified instances divided by total number of instances.

Majority class accuracy is the number of correct classified instances in specific majority class which is divided by means of total instances in particular majority class.

Minority class accuracy is amount of correctly classified instances in the particular minority class that gets divided in terms of total amount of instances in that minority class.

C. Results

TABLE 3: BALANCE DATASET - ACCURACY

Accuracy	Imbalance	RUS	ROS	SMOTE	UOS
Overall	94.56	86.24	89.88	86.53	**92.76**
Minority class	63.27	12.24	99.49	96.43	**97.96**
Majority class 1	96.88	**95.83**	84.03	81.25	89.24
Majority class 2	97.57	89.24	82.64	85.07	**90.97**

TABLE 3: NSL DATASET - ACCURACY

Accuracy	Imbalance	RUS	ROS	SMOTE	UOS
Overall	99.07	97.88	98.54	98.83	**98.87**
Minority class 1	98.70	96.19	98.09	**98.92**	98.88
Minority class 2	98.71	96.34	98.25	98.27	**98.38**
Majority class	99.45	99.56	99.29	99.29	**99.40**

Compared to RUS, ROS & SMOTE, UOS performs very well on three class imbalanced datasets. Out of 8 times, 6 times UOS outperforms other sampling methods.

TABLE 4: ABALONE DATASET - ACCURACY

Accuracy	Imbalance	RUS	ROS	SMOTE	UOS
Overall	95.06	70.50	65.80	74.92	**75.38**
Minority class	14.29	71.43	58.05	68.50	**71.99**
Majority class	100	69.04	73.44	**81.27**	78.57

TABLE 5: CENSUS DATASET - ACCURACY

Accuracy	Imbalance	RUS	ROS	SMOTE	UOS
Overall	97.69	89.52	88.17	90.16	**90.56**
Minority class	54.29	85.71	88.93	**96.25**	91.43
Majority class	100	**94.29**	87.52	86.45	89.65

TABLE 6: VOTE DATASET - ACCURACY

Accuracy	Imbalance	RUS	ROS	SMOTE	UOS
Overall	96.42	**96.12**	95.18	94.90	95.73
Minority class	88.24	86.76	93.38	95.13	**95.59**
Majority class	98.50	96.5	**97**	94.67	95.88

Comparing with RUS, ROS & the SMOTE, UOS could perform very good on the two class of imbalanced datasets. Out of 9, 4 times the UOS outperforms other specific sampling methods. 2 times the SMOTE, 2 times the RUS and 1 time that ROS which outperforms sampling techniques. In all case, accuracy of specific minority class is enhanced by comparing to other specific sampling method, during the usage of UOS sampling approach.

IV. CONCLUSION& FUTURE WORK

This review introduces the new approach for the imbalanced datasets in those two class and multi class environment with the increment the predicting accuracy of the minority class(es). This approach finds out the sampling value, and it is based on the sampling value thus performs under and over sampling. The result shows up the newly introduced sampling method enhances the specific prediction accuracy for minority class(es) including slight reduction of prediction accuracy of those majority class(es).

Future work can be considered more than three class environment and evaluate the performance of new sampling method. Combination of other sampling methods like SMOTE, borderline SMOTE, cluster based under sampling methods can be considered for UOS.Different algorithm level approaches without using data level approaches can be

considered for balancing the dataset. Especially improvement of AdaBoost for imbalanced datasets can be considered. Combination of UOS with algorithm level approaches can be considered to evaluate the performance on imbalanced datasets.

REFERENCES

[1]. Joffrey L. Leevy, Taghi M. Khoshgoftaar, Richard A. Bauder, NaeemSeliya, "A survey on addressing high-class imbalance in big data", Journal of Big Data, https://doi.org/10.1186/s40537-018-0151-6, 2018.

[2]. Yang Zhao,ZoieShui-Yee Wong, Kwok Leung Tsui, "A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection", Journal of Healthcare Engineering, 2018.

[3]. Jae-Hyun Seo and Yong-Hyuk Kim, "Machine-Learning Approach to Optimize SMOTE Ratio in Class Imbalance Dataset for Intrusion Detection", Computational Intelligence and Neuroscience, https://doi.org/10.1155/2018/9704672, 2018.

[4]. YangeSuna,b, ZhihaiWanga,HongtaoLia, Yao Lia, "A Novel Ensemble Classification for Data Streams with Class Imbalance and Concept Drift", International Journal of Performability Engineering, Vol. 13, No. 6, pp. 945-955, DOI: 10.23940/ijpe.17.06.p15.945955, 2017.

[5]. R.Barendela, R.M.Valdoyinos and J.S.Sanchez, "New Applications of ensembles of the classifiers. Pattern Analysis and Applications", (2003), 245-256.

[6]. G.E.A.P.A. Batista, R.C. Prati and M.C. Monard, "Study of the behavior of several methods for balancing the machine learning training data", SIGKDD 6:1 (2004), 20-29.

[7]. Chao Chen, Andy Liaw, Leo Breiman, "Using Random Forest to Learn Imbalanced Data", 2004.

[8]. N.V.Chawla, K.W.Bowyer, L.O.Hall and W.P. Kegelmyer, "SMOTE: Synthetic Minority Over–Sampling Technique", Journal of the Artificial Intelligence Research, 16, (2002).

[9]. N.V.Chawla, A. Lazarevic, L.O. Hall and K. Bowyer, "SMOTE Boost: Improving prediction of the Minority Class in the Boosting", 7th European Conference on the Principles and Practice of Knowledge Discovery in Databases, Cavtat (2003), 107-119.

[10]. Y. Freund and R.E. Schapire, "A decision - theoretic Generalization of the on-line learning and an application to boosting", Journal of Computer and System Sciences 55:1 (1997), 119-139.

[11]. H. Han, W.Y. Wang and B.H. Mao, "Borderline-SMOTE: A new over-sampling method in Imbalanced data sets learning", 2005 International Conference on Intelligent Computing (ICIC05).Springer 2005, Hefei (China, 2005) 878-887.

[12]. NiteshV.Chawla, Nathalie Japkowicz and AleksanKolcz, "Editorial: Special Issue on Learning from the Imbalanced Data Sets", SIGKDD Explorations 6 (1), (2004), 1-6.

[13]. C.Seiffert, T.Khoshgoftar and J.VanHulse, "Rus Boost: A Hybrid approach to alleviating the class imbalance", IEEE Transactions on Systems, Man and Cybernetics, Part a 40:1 (2010), 185-197.

[14]. S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models", IEEE Symposium, Computational Intelligence and Data Mining (IEEE 09). Nashville TN (USA, 2009), 324-331.

[15]. E.Prabhakar, M.Santhosh, A.Hari Krishnan, T.Kumar, R.Sudhakar, "Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Volume 7, Issue 01, April 2019.

[16]. E.Prabhakar, R.Parkavi, N.Sandhiya, M.Ambika, "Public Opinion Mining For Government Scheme Advertisement", International Journal of Information Research and Review, Volume 3, Issue 4, Page No.2112-2114, April 2016.

[17]. SeyyedaliFattahi, Zalinda Othman, Zulaiha Ali Othman, "New approach withensemble method to address class imbalance problem", Journal of Theoretical and Applied Information Technology, 2015.

[18]. D.Gopika, B.Azhagusundari, "A Novel Approach on Ensemble Classifiers withFast Rotation Forest Algorithm", International Journal of Innovative Research inComputer and Communication Engineering, Vol. 2, Issue 8, August 2014.

[19]. S. Yen and Y. Lee, "Under - sampling approaches for improving prediction of minority class in an imbalanced data set", International Conference on the Intelligent Computing (ICIC06), Kunming (China, 2006), 731-740.

[20]. E.Prabhakar, G.Pavithra, R.Sangeetha, G.Revathy, "Mining Better Advertisement Tool for Government Schemes", International Journal for Technological Research in Engineering, ISSN (Online): 2347 – 4718, Volume 3, Issue 5, Page No.1023-1026, January 2016.

[21]. Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano, "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, May 2011.

[22]. V.S.Suresh kumar, A.chandrasekar" *Fuzzy-GA Optimized Multi-Cloud Multi-Task Scheduler For Cloud Storage And Service Applications*" International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013.