Journal of Science Research

Improved data sampling schemes for alleviating class imbalance problem

Folorunso, S. O.* and **Adeyemo, A. B.****
*Department of Mathematical Sciences, Olabisi Onabanjo University, Ago Iwoye, Ogun State.
**Department of Computer Science, University of Ibadan, Ibadan, Oyo State.
Corresponding author: bamidelekeke@gmail.com

Abstract

The class imbalance problem occurs when standard classifiers are biased towards the majority class while the minority class is ignored. Existing classifiers tend to maximise overall prediction accuracy and minimise error at the expense of this minority class. However, studies had shown that misclassification cost of the minority class is higher and should not be ignored since it is the class of interest. This paper presents new improved data sampling schemes that can improve the classification performance of imbalance datasets and also increase the recall of the minority class. This paper also evaluates the performances of the improved schemes as well as the existing schemes using Receiver Operator's Characteristics (ROC) and recall of the minority class and Friedman Test for statistical analysis. This study was conducted using seven different base classifiers on three datasets from different domain to compare existing sampling techniques with the current. The improved sampling schemes often outperform the existing sampling schemes and is recommended for pre-processing of imbalance datasets before classification so as to improve classification performance and increase the recall of the minority class over the existing schemes.

Keywords: Imbalanced dataset; receiver operator's characteristics; ensemble learning; cost-sensitive learning.

Introduction

The class imbalance problem corresponds to the domain for which one class (majority) is represented by a large number of examples while the other (minority) is represented by few [1]. Sometimes, the ratio of these minority to majority classes can be as drastic as 1:100, 1:1,000, or 1: 10,000 and even more in some practical applications [2]. When a prediction model is trained on a dataset with such problem, it tends to show a strong bias towards the majority class, since typical classifiers (Decision Trees, Artificial Neural Networks, etc.) intend to maximize the overall prediction accuracy. Hence, the classification performance will be sub-optimal. However, the cost of mis-classifying the minority classis usually much higher than that of majority class and should not be ignored [3-5]. This is a fundamental problem of data mining research [6] and pattern recognition [7]. Some of the domains that suffers naturally from these class imbalances include intrusion detection [8], earth quakes, nuclear explosion and helicopter [9], risk management [2] text classification [10] education (the ratio of the number of "pass student" to "fail student") and detection of fraudulent or default banking [4].

Numerous existing solutions to Class Imbalance Problem were developed both at data and algorithm levels and reviewed at [11].

Examples of these solutions includeWilson's Edited Nearest Neighbour (ENN) Rule [12], Neighbourhood Cleaning Rule (NCL) [13], Tomek Links (TLink) [14],Condensed Nearest Neighbour (CNN) Rule [15], Random Under Sampling (RUS) [16], Synthetic Minority Oversampling Technique (SMOTE) [17], Random Over Sampling (ROS) [16], Cluster Based Over Sampling (CBOS) [18], One Sided Selection (OSS) [19], SMOTE + ENN, SMOTE + Tlink and CNN+ TL [20].

Examples of some solutions developed at the algorithm level were either by adjusting the algorithm itself [21], One-Class Learning [2], Cost Sensitive and Ensemble Learning [22].

Related work

Many works that had addressed the class-imbalance problem includes Hulse *et al* [16], Kubat and Matwin [19] and [20] Batista *et al* where sampling schemes (RUS, ROS, OSS, CBOS, ENN, SMOTE) were used to alleviate class imbalance problem on some public data set.

Also, Habibi *et al* [23], Awokola [24] and Agboola [25] used these data mining techniques (Decision Tree, ANN, REP Tree, RIPPER, Ridor rules, Random Tree, Decision Stump, CART, LADTree, NB and Functional Tree) directly on data sets without using any sampling schemes. Pair *t*-Test and correlation statistical test were used with ROC, RMSE and Accuracy as metrics. However, all the datasets used were highly skewed thus the classification performance were suboptimal. Batista *et al* [20] proposed using CNN+TLink, SMOTE + TLink and SMOTE + ENN schemes on a two-class problem with Decision Tree classifier (both Pruned and non-Pruned) with Hsu's Multiple Comparison with the Best (MCB), ROC and Geometric mean of 13 different datasets from UCI data repository. Lessmann [21] adjusted SVM's internal parameterization to classify Customer Response management data. He used ROS and RUS sampling scheme with *f*-Measure and Geometric mean as the performance metric. But the SVM classifier used cannot be reused for another data-set.

However, this study presents new improved data sampling schemes that can improve the classification performance of imbalance data-sets and also increase the recall of the minority class.

The rest of the text thus outlined. Section 2 includes the methodology adopted for the research, the data sets used and the metrics used for performance evaluation. Section 3 presents the result obtained and also discussed the result. Finally, the conclusion and future work.

Materials and methods

The problem identified was that; for an extremely skewed class distribution, the recall of the minority class is often 0, which means that there are no classification rules generated for the minority class.

Data-sets

The three data-sets used for the study were:

1. Diabetes Mellitus (DM) disease data-set which was obtained from Wesley Guilds Hospital, Ilesa by Awokola [24]. The data-set contained 886 instances of complete record of DM patients from January 2009 to May 2010.
2. Senior Secondary School (SSS) result examination result data-set obtained from the West African Examination Council (WAEC) Office in Ibadan by Agboola [25]. The data-set spans a period of five years (2005-2009).
3. Contraceptive Methods (CM) dataset obtained from the Health Centre of Ibadan North East Local Government, Ibadan [26]. The data covers a period of seven (7) years (2008-2014). The research datasets were summarized in Table 1.

Table 1. Summary of research data-set.

Data-sets	No. of instances	Attributes	No. of classes	% minority class
DM	886	19	3	2
SSS Result	1163	8	4	4
CM	775	20	5	7

The data-sets were pre-processed with the existing (ENN, NCL, CNN, SMOTE, RUS, SMOTE300 and SMOTENN) and improved (SMOTE300ENN, SMOTE300RUS, SMOTENCL, SMOTERUS and SMOTE300NCL) data sampling schemes. The improved data sampling schemes were coded using Java programming language and implemented in Waikato Environment for Knowledge Analysis (WEKA) version 3.6.8, an open source data mining suite [27] as additional filters available for use in the filter library.

New data sampling scheme algorithm

Given a dataset, *T*, which consists of pair (x_i, y_i), where $i = 1, 2, \ldots m$ where x_i denote the input attributes and denote the class labels. *T* contains *n* instances with *m* attributes each and either belong to a positive (minority) or negative (majority) class. The minority Class *C* which is also the class of interest is a subset of $y_i : C \subset y_i$. The minority class was oversampled by taking each minority class samples and introducing synthetic examples along the line segments joining any/all of the *k* nearest Neighbours.

Algorithm SMOTE (*T, N, k*)

Input: Number of minority class samples *T*; amount of

SMOTE N%; Number of Nearest Neighbors, *k*.
Output: (N/100) **T* synthetic minority class samples.

1. (* if *N* is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd.*).
2. If *N* <100.
3. Then Randomize the T minority class samples.
4. *T* = (N/100) * *T*.
5. *N* = 100.
6. End if:
7. *N* = (int) (N/100) (*The amount of SMOTE is assumed to be in integral multiples of 100*).
8. *k* = Number of Nearest Neighbhour.
9. numattrs = number of attributes.
10. Sample [ ] [ ]: array for original minority class samples.
11. new index: keeps a count of number of synthetic samples generated, initialized to 0.
12. Synthetic [ ] [ ]: array of synthetic samples.
    (* compute *k* nearest neighbors for each minority class sample only*).
13. for *i* ¬ 1 to *T*.
14. compute *k* nearest neighbors for *i,* and save the indices in the nnarray.
15. Populate (*N*, *i*, *nn*array).
16. end for:
    Populate (*N*, *i*, *nn*array) (*Function to generate the synthetic samples*).
17. While *N* 0 do.
18. Choose a random number between 1 and *k*, call it *nn*. This step chooses one of the *k* nearest neighbor of *i*.
19. for attr ← 1 to numattrs.
20. If attr = continuous feature.
21. Compute: *di f* = Sample [nnarray [nn]] [attr] – Sample [*i*] [attr].
22. Compute: *gap* = random number between 0 and 1.
23. Synthetic[newindex] [attr] = Sample [*i*] [attr] + gap * *dif*.
24. else:
25. attr_value = majority vote for the attr values between *i* and *nn*. If no majority, then choose at random.
26. synthetic [newindex] [attr] = attr_value.
27. end for;
28. new index ++.
29. *N* = *N* – 1
30. end while:
31. return (* End of pseudo-code for SMOTE*).

In this study, *k* = 5-Nearest Neighbour was used and the rate of over-sampling used was 300%. The value of *k* is arrived at after some series test from 100-500%. This over-sampling caused the selection of a random point along the line segment between three specific features as against the original one specific feature. This approach effectively forced the decision region of the minority class to become more general. Then, ENN, NCL and RUS data sampling schemes were applied respectively to both original SMOTE and SMOTE+300% to remove noisy and erroneous data in the dataset.

Evaluation metrics

Confusion matrix is a table that records the result of correctly and incorrectly recognised examples of each class [2] and is presented in Table 3. The positive class represents the minority class while the negative class represent the majority class. True Positive (TP) shows the number of positive class correctly classified as positive, while True Negative (TN) shows the number of negative class correctly classified as negative class. False Positive (FP) shows the number of negative classes that were incorrectly classified as the positive class while false negative (FN) shows the number of positive classes that were incorrectly classified as negative class. The Recall is the likelihood that a positive class is correctly classified as positive as depicted by equation (1). Equations 1, 2 and 3 are all derived from Table 2.

Table 2. Confusion matrix.

	Positive Prediction	Negative Prediction
Positive Class	TP	FN
Negative Class	FP	TN

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad \ldots\ldots (1)$$

$$\text{FPR} = \frac{\text{FP}}{(\text{FP}+\text{TN})} \quad \ldots\ldots (2)$$

$$\text{ROC} = \frac{1+(\text{TPR}-\text{FPR})}{2} \quad \ldots\ldots (3)$$

Result

A total of thirteen balanced data-sets which were created from thirteen different data-sampling schemes (both existing and improved) were trained on seven different base classifiers. The results obtained were analyzed statistically using Friedman test with ROC and Recall of the minority class as performance metrics. The higher the mean rank value, the better the scheme.

Report of Friedman's test on recall of the minority class metric for all datasets

The results obtained with Friedman's test on recall metric with their mean rank values is presented in Table 3. This analysis showed that SMOTE300ENN scheme had the highest mean rank values of 12.04, 13.00 and 12.64 respectively across the three data-sets. Thus, Friedman test established that SMOTE300ENN scheme, one of the improved data sampling schemes gave the best recall value. This means that the minority class across the three datasets were well detected with this scheme. This analysis showed that ENN scheme had lowest mean rank values of 1.64 and 2.04 for DM and SSS Result data-sets respectively and NCL scheme had the lowest mean rank value of 1.93 for CM data-set. These schemes gave zero or no detection of the minority class.

It was remarkable to mention that consistently, four of the improved data sampling schemes (SMOTE300ENN, SMOTE300RUS, SMOTE300NCL and SMOTERUS) were generally ranked amongst the best seven out of all the data-sampling schemes.

Table 3. Report of Friedmans test on recall of Minority class metric for all data-sets.

S/N	Data sampling schemes	DM	SSS result	CM
1	CNN	2.68	2.79	2.07
2	ENN	1.64	2.04	3.46
3	NCL	4.32	3.64	1.93
4	RAWDATA	2.29	2.11	3.54
5	RUS	6.54	7.86	5.43
6	SMOTE	5.04	5.29	6.46
7	SMOTE300	7.29	8.54	9.89
8	SMOTE300ENN	12.04	13.00	12.64
9	SMOTE300NCL	9.82	8.57	10.61
10	SMOTE300RUS	10.68	11.32	12.14
11	SMOTEENN	10.54	10.79	7.54
12	SMOTENCL	8.36	5.36	6.32
13	SMOTERUS	9.79	9.71	8.54

Report of Friedman's test on ROC metric for all classifiers

The results obtained with Friedman's test using ROC metric on seven different classifiers with their mean rank values are presented in Table 4. The analysis revealed that: Decision Tree gave the best classification performance on all data sampling schemes on DM and SSS Result dataset while SVM surpassed the other data sampling schemes with best classification ability on CM dataset.

For all the classifiers considered in this study, their ROC values were greater than 0.5. Though, least performances were recorded for SVM and ANN across all data-sets, they were still far from random guessing. Japkowicz [28] and Carvajal *et al* [29] agreed that the reason for deficient performance of ANN classifier was due to the fact that minority class was inadequately weighted in networks. Bhatnagar [30] and Batuwita [31] agreed that though SVM can handle class imbalance problem, it got overwhelmed when faced with more severe class imbalance problem.

Table 4. Report of Friedmans test on ROC metric for all classifiers.

S/N	CLASSIFIER	DM	SSS Result	CM
1	SVM	1.50	2.64	6.71
2	RIPPER	3.61	1.93	2.54
3	REPTREE	5.54	5.36	1.86
4	RANDOMTREE	2.71	4.61	4.07
5	ANN	4.71	2.89	2.96
6	KNN	3.68	5.07	5.50
7	DECISIONTREE	6.25	5.50	4.36

Discussion

From the analysis result, it could be deduced that CNN scheme performed least across all the three datasets. One of the reasons for this could be that the scheme does not guarantee a minimal subset as an under-sampling scheme [32]. CNN scheme dropped over 50% of the data. This led to loss of information for a classifier to work with. Moreover, it is especially sensitive to noise as the scheme only removes redundant examples that are far from the decision border from the dataset thereby retaining noisy instances. This corroborates the report of Dasarathy *et al* [33]. Hence, this data sampling scheme is good when memory requirement [34] and computational advantage [33] is the main concern.

The ENN scheme 'clean' the possible overlapping (border) region of the different classes, leaving smoother decision boundaries as corroborated by [35]. However, the minority class was still ignored and not detected.

RUS scheme often performed better than CNN scheme across all three datasets. Though instances were removed randomly from the data-set to give a balanced distribution i.e. the size of all the classes were the same, the scheme gave a better recognition or increases the class bias of the minority class as corroborated by [16].

NCL scheme's good performance could be due to the fact that it also 'cleans' the data-set before classification like its predecessor (ENN) but only the majority class. It also 'cleaned' neighbourhood that misclassifies examples belonging to the minority class which is the class of interest as corroborated by [13]. SMOTE scheme increases the size of the minority class; so also, will the class sub-clusters and boundary points increase. This scheme synthetically increases the number of the minority class which is also the class of interest. But the detection of the minority class increased and also their decision boundary as supported by [36].

SMOTE300ENN scheme performed best out of all the schemes. One of the reasons for this could be that when SMOTE300 scheme was applied to the original data-set (RAWDATA), it increased the size of the minority class by 300% for better recognition. But this did not solve the problem of different class clusters. In order to create better-defined class clusters, ENN scheme was applied to the dataset created from SMOTE300 scheme to remove noisy, erroneous and mis-classified instances from other classes. Hence, this scheme provides a set of instances organised in relatively compact and homogeneous sub-group for better detection of both majority and minority classes for optimal classification. It also solved the problem of class overlapping and sub-class clusters. The new data-set created from this scheme is free from noise, errors and class overlap.

The advantage of SMOTE300RUS data-sampling scheme was that all class probability and the sizes of the classes were the same. This data sampling scheme further enhanced the decision region of the minority class and better detection. But the noise level is still as in RAW DATA. This scheme may not be recommended for a data-set with highly overlapped-classes.

SMOTE300NCL performed well also as the application of NCL removed noisy, erroneous and mis-classified instances from only the majority class instances while the size of the minority class remains the same. With the reduced data-set, it was difficult to maintain the original classification accuracy. The new data-set created from SMOTE300NCL data-sampling scheme was free from noise from the majority but not from the minority class.

SMOTERUS scheme is similar to SMOTE300RUS. SMOTENCL scheme is also similar to SMOTE300NCL.The new dataset created from the application of SMOTENCL will be free from noise from the majority-class but not from the minority-class.

Conclusion

This study reviewed the Class Imbalance Problem and some of the existing solutions and domains where class imbalance problem occurred. Results showed that new improved data sampling schemes increased the recall of the minority class across all data-sets when compared to the original data-sets. The improved data-sampling schemes can be applied to highly skewed data-sets with a very small number of minority classes as they perform well in the detection of the minority class. Future work proposed is to add Cost Sensitive Learning (CSL) to the new improved schemes.

References

[1] Japkowicz, N. 2003. Class Imbalances: Are we focusing on the right issue? In: *The Proceeding of ICML Workshop: Learning with Imbalanced Data Sets II.*

[2] Chawla, N. V., Japkowicz, N. and Kolcz, A. 2004. Editorial: Special Issue on Learning from Imbalance Data Sets. *SIGKDD Explorations,* Vol. 6, No. 1, pp. 1-6.

[3] Hido, S. and Kashima, H. 2008. Roughly Ballanced Baggingfor Imbalanced Data, in *SIAM Conference on Data Mining (SDM2008)*, Atlanta, Georgia.

[4] Thai-Nghe, N. Andre, B. and Lars, S. 2009. Improving Academic Performance Prediction by Dealing with Class Imbalance, in *9th IEEE International Conference on Intelligent Systems Design and Applications,* pp. 878-883.

[5] Napierala, K. and Stefanowski, J. 2016. Types of minority class examples and their influence. *Intelligent Information System,* Vol. 46, p. 563-597.

[6] Yang, Q. and Wu, X. 2005. 10 challenging problems in data mining. *International Journal of Information Technology and Decision making,* Vol. 5, No. 4, pp. 597-604.

[7] Ghanem, A., Venkatesh, S. and West, G. 2010. Multi-Class pattern classification in imbalanced data. In: *Pattern Recognition,* pp. 2881-2884.

[8] Vegard, E. 2010. Machine learning for network based intrusion detection.

[9] Guo, X., Yin, Y., Dong, C., Yang, G. and Zhou, G. 2008. On the class imbalance problem. In: *Fourth International Conference on Natural Computation, IEEE Computer Society,* pp. 192-201.

[10] Estabrooks, A., Jo, T. and Japkowicz, N. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence,* Vol. 20, No. 1, pp. 18-36.

[11] Folorunso, S. O. and Adeyemo, A. B. 2012. Theoretical comparison of undersampling techniques.

Proceedings of the 2nd International Conference Computing, Energy, Networking, Robotics and Telecommunications (EIECON).

[12] Wilson, D. L. 1972. Asymptotic properties of nearest neighbour rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 2, No. 3, pp. 408-421.

[13] Laurikala, J. 2001. Improving identification of difficult small classes by balancing class distribution. *Proceedings of the International Conference on Artificial Intelligence in Medicine in Europe*, Vol. 2101, No. 8, pp. 63-66.

[14] Tomek, I. 1976. Two modification of CNN. *IEEE Transactions on Systems, Man and Communications, SMC,* Vol. 6, pp. 769-772.

[15] Hart, P. E. 1968. The condensed nearest neighbour rule. *IEEE Transformation Information Theory,* Vol. 16, pp. 515-516.

[16] Hulse, J. V., Khoshgoftaar, T. M. and Napolitano, A. 2007. Experimental perspectives on learning from imbalanced data. *24th International Conference on Machine Learning (ICML),* pp. 935-942.

[17] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, P. W. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research (JAIR),* Vol. 16, pp. 321-357.

[18] Jo, T. and Japkowicz, N. 2004. Class imbalances versus small disjuncts. *SIGKDD Explorations* , Vol. 6, pp. 40-49.

[19] Kubat, M. and Matwin, S. 1997. Addressing the curse imbalanced training sets: One-sided selection. *Proceedings of the Fourteenth Conference on Machine Learning,* pp. 179-186

[20] Batista, G. E. A., Prati, R. C. and Monard, M. C. 2004. A study of the behaviour of several methods for balancing machine learning training data. *SIGKDD Explorations,* Vol. 6, No. 1, pp. 20-29.

[21] Lessmann, S. 2004. Solving imbalanced classification problems with support vector machines. *International Conference on Artificial Intelligence,* pp. 214-220.

[22] Sun, Y., Kamel, M. S. and Wang, Y. 2006. Boosting for learning multiple classes with imbalanced class distribution. *6th Internation Conference on Data Mining (ICDM),* pp. 7695-7701

[23] Habibi, S., Ahmadi, M and Alizadeh, S. 2015. TYPE 2 diabetes mellitus screening and risk factors using decision tree: Result of data mining, *Global Journal of Health Sciences,* Vol. 7, No. 5, pp. 304-310.

[24] Awokola, E. O. 2010. Rapid diagnosis and treatment of diabetes mellitus using artificial neural networks. University of Ibadan, Ibadan.

[25] Agboola, L. A. 2010. Development of a decision tree based model for analysing student performance in two national certificate examinations (WAEC and NECO). University of Ibadan, Ibadan.

[26] Folorunso, S. O. 2015. Development of advanced sampling schemes to alleviate class imbalance problem in data mining classification algorithms. University of Ibadan, Ibadan.

[27] Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. A. 2010. WEKA – Experiences with a Java open-source project. *Journal of Machine Learning Research,* Vol. 11, pp. 2533-2541.

[28] Japkowicz, N. and Stephen, S. 2002. The class imbalance problem: A aystemic study. *IDA Journal,* Vol. 6, No. 5, p. 429-449.

[29] Carvajal, K., Chacon, M., Mery, D. and Acuna, G. 2004. Neural network method for failure detection with skewed class distribution. *Journal of the British Institute of Non-Destructive Testing,* Vol. 46, No. 7, pp. 399-402.

[30] Bhatnagar, V., Bhardwat, M. and Mahabal, A. 2010. Comparing SVM ensembles for imbalanced datasets. *10th IEEE International Conference on Intelligent Systems Design and Applications (ISDA),* pp. 651-657.

[31] Batuwita, R. and Palade, V. 2012. *Imbalanced Learning: Foundations, Algorithm and Applications*, H. Haibo and M. Yunqian, Eds., John Wiley and Sons Inc.

[32] Wilson, D. R. and Martinez, T. R. 2012. An Integrated instance-based learning algorithm. *Computational Intelligence,* Vol. 16, No. 1, pp. 1-28.

[33] Dasarathy, B. V., Sanchez, J. S. and Townsend, S. 2000. nearest neighbour editing and condensing tools – synergy exploitation. *Pattern Analysis and Applications,* Vol. 3, pp. 19-30.

[34] Bathia, N. and Vandana, A. 2010. Survey of nearest neighbour techniques. *International Journal of Computer Science and Information Security (IJCSIS),* Vol. 8, No. 2, pp. 302-305.

[35] Vazquez, F., Sanchez, J. and Pla, F. 2005. Pattern recognition and image analysis. in *Second Iberian Conference ibPRIA 20005, Part II,* pp. 35-42.

[36] Dittman, D. J., Taghi, M. K., Randall, W. and Amri. 2014. *The Proceedings of the 27th International Florida Artificial Intelligence Research Society Conferenc*e.

Journal of Science Research
Volume 15, 2016, ISSN 1119 7333
Citation: Folorunso, S. O. and **Adeyemo, A. B.**
Improved data sampling schemes for alleviating class imbalance problem.

Textflow Limited
Ibadan, Nigeria