




# Cluster-Based Under-Sampling Using Farthest Neighbour Technique for Imbalanced Datasets

G. Rekha<sup>1</sup>(✉) and Amit Kumar Tyagi<sup>2</sup>(✉) 

<sup>1</sup> Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India

[gillala.rekha@klh.edu.in](mailto:gillala.rekha@klh.edu.in)

<sup>2</sup> School of Computing Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai 600127, Tamilnadu, India

[amitkrtyagi025@gmail.com](mailto:amitkrtyagi025@gmail.com)

**Abstract.** In domain of data mining, learning from imbalanced class distribution datasets is a challenging problem for conventional classifiers. The class imbalance exists when the number of samples of one class is much lesser than the ones of the other classes. In real-world classification problems, data samples often have unequal class distribution. This problem is represented as a class imbalance problem. However, many solutions have been proposed in the literature to improve classifier performance. But recent works entitlement that imbalanced dataset is not a problem in itself. The degradation of classifier performance is also linked with many factors like small sample size, sample overlapping, class disjunct and many more. In this work, we proposed cluster-based under-sampling based on farthest neighbors. The majority class samples are selected based on the average distance to all minority class samples in the cluster are farthest. The experimental results show that our cluster-based under-sampling approach outperform with existing techniques in the previous studies.

**Keywords:** Classification · Clustering · Class disjunct · Imbalance problems · Majority samples · Minority samples

## 1 Introduction

In the current big data era, data mining and machine learning play a vital role in effective decision making. Among that classification is one of the important techniques most widely used for various application from healthcare to a business decision such as bankruptcy prediction [1], cancer prediction [2], churn prediction [3], face detection [4], fraud detection [5], and software fault prediction [6]. In general, the performance of the classifier is associated with data distribution. If equal or balanced distribution of data will increase the performance of the classifier. However, most of the real world is usually skewed in nature, i.e., the number of samples from one class will be more than the other class. For example, for binary classification, if one class has 1000 samples

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

A. Abraham et al. (Eds.): IBICA 2019, AISC 1180, pp. 35–44, 2021.

[https://doi.org/10.1007/978-3-030-49339-4\\_5](https://doi.org/10.1007/978-3-030-49339-4_5)

(majority samples/negative samples) and the other class has 100 samples (minority samples/positive samples) will lead to bias when trained on classification algorithm. The skewed distribution of data is not a problem in itself. Apart, small sample size, small disjunct, sample overlapping, etc., will degrade the performance of the classifier.

To address the problem of class imbalance, the machine learning community has relied upon three techniques in general consist of data pre-processing technique, algorithm-level technique and ensemble learning technique.

- **Data pre-processing:** In these techniques, the skewed data is balanced prior to training on a classifier. These techniques are simple and easy to implement. The most popular data pre-processing technique is sampling. The sampling method consists of over-sampling and under-sampling techniques. In oversampling, synthetic data samples are generated for minority samples to balance the distribution. Some of the oversampling techniques are Random OverSampling (ROS), Synthetic Minority Oversampling Technique (SMOTE). For under sampling, the majority samples are discarded to balance the distribution. Some of the techniques are Random Under Sampling (RUS), Tomek Link.
- **Algorithm-level:** In these techniques, the existing algorithms are modified by applying or adjusting the weights or including of loss function. It implements within the algorithm itself during the learner phase.
- **Ensemble or Hybrid level:** In these techniques, the combination of data level and algorithm level techniques are used to provide solutions for class imbalance problems. It allows multiple classifiers to be modeled at the same time and create a final model to generate better accuracy.

Hence, the organization of this paper is follows as. Section 2 discusses the related work in imbalance classification problem. Section 3 presents the methodology of the proposed work. Experiments and Results are discussed in Sect. 4. Section 5 presents concluding remarks.

## 2 Related Work

As specified in the introduction, the class imbalance problem is crucial and an effective solution is much in demand. Since traditional classification algorithms are not designed to be trained on imbalanced datasets. It basically leads to a series of problems, overfitting of majority/negative classes and under fitting of positive/minority classes. Apart from imbalanced nature, the other problems like small subsamples, overlapping of samples, small disjunct, noise occur in the dataset. To handle class imbalance problems, broadly three methods are been proposed in literature a. Data Level methods b. Algorithm-level methods and c. Ensemble methods.

In data-level methods, data resampling is performed to balance the distribution of data before training on a classifier. In algorithm-level methods, the traditional classification algorithms are modified to handle imbalanced data either by adjusting cost or weights. The third technique is ensemble methods where multiple classifiers are trained and the majority voting method is used to select the best classifier.

As suggested in the literature, rebalancing the datasets at the data-level is simple and effective to avoid the bias in classification [27]. It is a pre-processing technique used to balance the data before training on a classifier. The common sampling methods used to balance the skewed distribution are Oversampling and under-sampling techniques. In oversampling, the minority samples/positive samples are resampled to generate synthetic data to meet the size of the majority class whereas, in under-sampling, the majority samples are discarded to meet the size of the minority class. The major drawback of the former is duplication of data generation and for the latter loss of important information. To overcome the drawback of oversampling, the Synthetic Minority Oversampling Technique (SMOTE) [9] has been proposed in the literature. It is one of the most popular and efficient technique. SMOTE generates synthetic data from its minority class sample neighbor using Euclidean distance. But, the major drawback is that the synthetic samples may overlap with its surrounded majority samples. To address this particular weakness may extend the version of SMOTE has been proposed by the research community, for example, Borderline SMOTE [10], MSMOTE [11] and etc. on the other hand, the under sampling technique may discard important or representative samples from the datasets. Kubat et al. [12] adopted one side selection to under-sampling the majority class by removing noise, boundary and redundant samples. Estabrooks et al. [13] proposed over-sampling and under-sampling techniques with different sampling rates to generate many sub-classifiers and finally integrated it. The results showed a better performance compared to ensemble methods.

In Algorithm-level, the class imbalance is addressed by directly modifying the classification algorithm or using different misclassified costs. These methods depend on the classifier to enhance classifier performance. Wu and Chang [14] proposed a method called Kernel-Boundary Alignment (KBA). KBA is based on Radial Basis Function (RBF) to compute the distance between all data points and also for class distribution. In [15], the author proposed a Confusion Matrix based Kernel LOGistic Regression (CM-KLOGR) for handling class imbalanced datasets. CM-KLOGR applied weighted harmonic mean to measure the performance metrics from the cost matrix. In recent years, Deep learning has become a popular research topic for feature representation. Khan et al. [16] proposed a Cost-Sensitive (CoSen) deep neural network to learn feature representations focused on image datasets. Dong et al. [17] proposed incremental minority class discrimination using a multi-label classification problem in deep learning to address class imbalance problem.

The ensemble techniques are popular techniques in handling class imbalance problems. They work by learning on multiple base classifiers and then it adopts an ensemble technique to improve the performance of the classifier. The most popular and frequently used ensemble approaches are bagging (bootstrap aggregating) [18], stacking [19] and boosting [20] techniques. Several researchers devised novel approaches that combine either oversampling or under sampling techniques into the ensemble framework. The variation of boosting includes SMOTEBoost [21], RUSBoost [22], DataBoost-IM [23] algorithm. The SMOTEBoost algorithm is an integration of SMOTE technique with AdaBoost algorithm and RUSBoost is an integration of RUS technique with AdaBoost algorithm. The DataBoost-IM algorithm combines AdaBoost with Gaussian distribution to generate synthetic samples. Rekha et al. [28] proposed a noise filtering approach to

remove the noise samples from the dataset and compared the performance using boosting and bagging techniques with and without noise filtering.

The bagging approach is similar to the boosting approach but bagging implements several sub-classifiers and selects the best classifier based on majority voting. The variations of bagging include UnderBagging (UB) [24], SMOTEBagging [25] and many more. UnderBagging adopts under-sampling to discard the majority samples and SMOTEBagging integrates smote with bagging techniques. Galar et al. [26] have provided a systematic review of ensemble techniques for class imbalance problem. They proposed a taxonomy for boosting, bagging approaches.

In the past lot of research work is focused on under-sampling the majority class sampling, oversampling the minority samples or tuning the parameters at the algorithm. But it is very important to recognize the majority of samples that are not overlapped with minority samples. To avoid the loss of important information while under sampling the majority samples, it is better to pick the majority samples based on its existence with that of minority samples. Hence, this section discusses about the related work for handling skewed distribution of data. Now, next section will deal with the proposed methodology using cluster-based techniques in addressing class imbalance problem.

### 3 Proposed Methodology

Majority samples are selected based on how well-suited the majority samples are with the minority samples by keeping the minority samples as a whole. In class imbalance problems, the majority/negative samples are overwhelmed the minority class samples and it's important to find the bias majority class samples as it may suffer the classification accuracy. Over-sampling techniques may crowd the minority samples by generating synthetic data but may cause overlapping between the samples. Under-sampling may discard the majority samples to meet the size of the minority samples. Some under-sampling techniques adopted clustering and instance selection methods. The major concern of under-sampling is how to reduce the majority samples in an effective way. In our proposed work, the majority class samples are selected based on the average distance to all minority class samples in the cluster are farthest.

In Fig. 1, the data points are represented using clusters and in each cluster, we calculated the distance between the majority data point with all the minority class samples. Based on the farthest average distance majority samples have been picked. In our work, we applied a Euclidean distance formula to calculate the distance between data points.

Figure 2 shows the proposed model. The entire data is been grouped using the k-means clustering approach. The value of k varies from 3 to 5 to check the best partition of data into different clusters. Once the clustering process is done, the selection of majority samples has been performed as mentioned above. Next, the selection majority samples are combined with minority samples to perform classification. Finally, the classifier is tested based on the test data to check the accuracy of the model. Hence, this section discusses proposed methodology in detail. Now, next section deal with the experimental and simulation results for the proposed methodology.

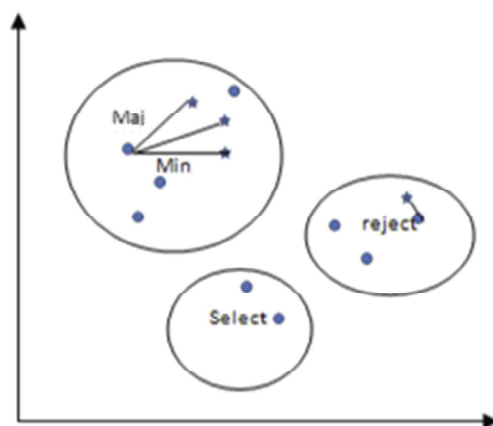


Fig. 1. Representation of majority and minority samples

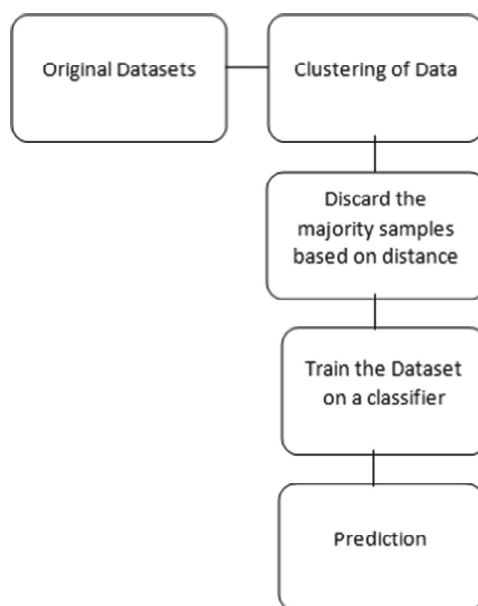


Fig. 2. The proposed model

## 4 Experimental and Simulation Results

In this section, we conducted the experiment on 10 datasets from keel repository<sup>1</sup>. The whole experiment is tested and verified using a 10-fold cross-validation technique. Table 1 shows the datasets used in the experiment. All the experiment has been run using the Decision tree algorithm.

<sup>1</sup> <https://sci2s.ugr.es/keel/imbalanced.php>.

**Table 1.** Datasets with their characteristics

Datasets	Size	# attr	% IR
Ecoli	336	7	3.36
Glass	214	9	6.38
Haberman	306	3	2.78
Iris	150	4	2
New-thyroid	215	5	5.14
Pima	768	8	1.87
Satimage	6435	36	9.28
Shuttle	1829	9	13.87
Vehicle	846	18	3.25
Wisconsin	683	9	1.86
Ionosphere	351	34	1.79

The evaluation criteria for imbalance problems are considered based on the confusion matrix. The different formulas for evaluation metrics are provided in Eqs. 1 to 5.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall/Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

$$\text{F-Measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

$$\text{G-Mean} = \sqrt{\left(\frac{TP}{TP + FN}\right) + \left(\frac{TN}{TN + FP}\right)} \quad (5)$$

In this paper, the performance of the proposed method is investigated on two evaluations metric such as F-Measure and G-Mean.

#### 4.1 Results

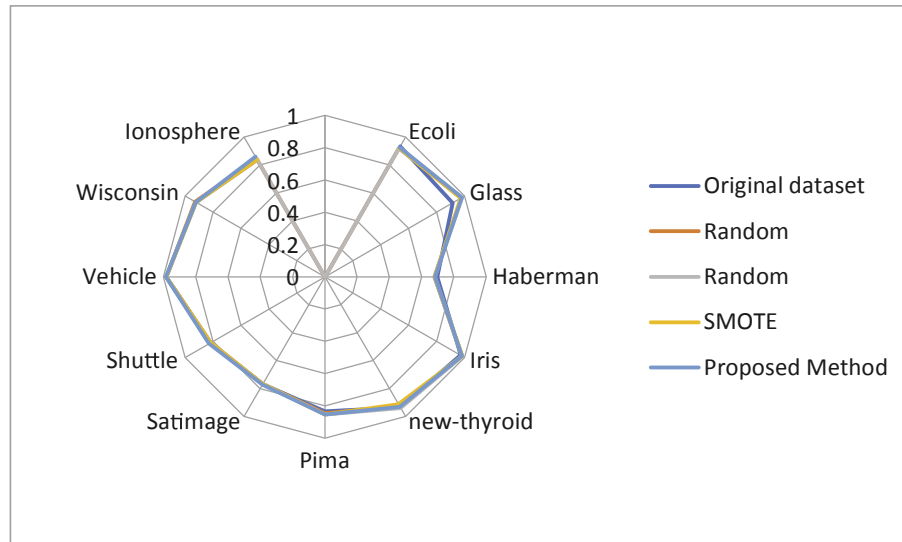
We applied the proposed method on 10 datasets and used F-Measure and G-Mean as the performance metrics. In the experiments, we had trained the model using original data with any sampling technique. Additionally, we performed random oversampling, under-sampling and SMOTE on the data. All four methods are used to compare our proposed model. The experiments are carried on the Decision tree algorithm (C4.5) with 10-fold cross-validation. The experimental results are presented in Table 2 and 3 and the

**Table 2.** F-Measure performance results

Data set	Original dataset	Random oversampling	Random under-sampling	SMOTE	Proposed method
Ecoli	0.9234	0.9197	0.9191	0.9232	0.9269
Glass	0.9456	0.9717	0.9911	0.9832	0.9912
Haberman	0.6795	0.6805	0.6800	0.6824	0.6838
Iris	0.9812	0.9832	0.9842	0.9812	0.9811
New-thyroid	0.9234	0.9197	0.9191	0.9232	0.9269
Pima	0.8245	0.8241	0.8225	0.8238	0.8248
Satimage	0.7618	0.7611	0.7610	0.7613	0.7625
Shuttle	0.8677	0.8695	0.8698	0.8684	0.8688
Vehicle	0.9732	0.9736	0.9702	0.9835	0.9809
Wisconsin	0.9132	0.9182	0.9210	0.9213	0.9234
Ionosphere	0.8386	0.8420	0.8111	0.8372	0.8511

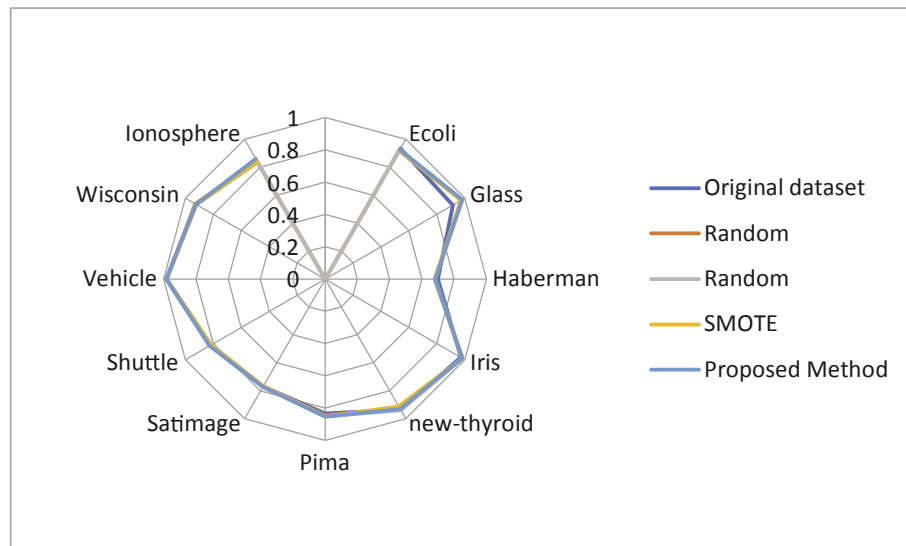
graphical representation of F-measure and G-mean performance is presented in Fig. 3 and 4. Among the 10 datasets, the proposed model achieved the best performance on 8 datasets.

Hence, this section presents experimental and simulation results of our proposed model. In which we provide some valid and efficient results. Now, next section will conclude this work in brief by providing some future enhancements.

**Fig. 3.** F-Measure performance

**Table 3.** G-Mean performance results

Data set	Original dataset	Random oversampling	Random under-sampling	SMOTE	Proposed method
Ecoli	0.9332	0.9197	0.9196	0.9232	0.9269
Glass	0.9145	0.9771	0.9819	0.9723	0.9819
Haberman	0.6982	0.6805	0.6800	0.6823	0.6838
Iris	0.9718	0.9820	0.9842	0.9821	0.9811
New-thyroid	0.9342	0.9291	0.9395	0.9132	0.9299
Pima	0.8342	0.8453	0.8532	0.8538	0.8548
Satimage	0.7698	0.7691	0.7690	0.7693	0.7725
Shuttle	0.8167	0.8169	0.8169	0.8168	0.8288
Vehicle	0.9834	0.9836	0.9820	0.9835	0.9849
Wisconsin	0.9232	0.9282	0.9220	0.9223	0.9234
Ionosphere	0.8386	0.8421	0.8432	0.8372	0.8589

**Fig. 4.** G-Mean performance

## 5 Conclusion

In real-world classification problems, imbalanced data occurred in different application domains and received considerable attention from the research community. The degradation of classifier performance is also linked with many factors like small sample size, sample overlapping, class disjunct and many more. In this work, we proposed



cluster-based under-sampling based on farthest neighbors. The majority class samples are selected based on the average distance to all minority class samples in the cluster are farthest. The experiment results indicate that the proposed work outperforms other sampling methods. Hence, in future we want to perform the experiments using ensemble classification algorithms and also apply for real time datasets.

**Acknowledgment.** This Research is funded by Anumit Academy's Research and Innovation Network (AARIN), India. The Author Would Like to Thank AARIN, India, a Research Network for Supporting The Project Through its Financial Assistance.

## References

1. Lin, W.-Y., Hu, Y.-H., Tsai, C.-F.: Machine learning in financial crisis prediction: a survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **42**(4), 421–436 (2012)
2. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotech. J.* **13**, 8–17 (2015)
3. Mahajan, V., Misra, R., Mahajan, R.: Review of data mining techniques for churn prediction in telecom. *J. Inf. Organ. Sci.* **39**(2), 183–197 (2015)
4. Zafeiriou, S., Zhang, C., Zhang, Z.: A survey on face detection in the wild: past, present and future. *Comput. Vis. Image Underst.* **138**, 1–24 (2015)
5. West, J., Bhattacharya, M.: Intelligent financial fraud detection: a comprehensive review. *Comput. Secur.* **57**, 47–66 (2016)
6. Malhotra, R.: A systematic review of machine learning techniques for software fault prediction. *Appl. Soft Comput.* **27**, 504–518 (2015)
7. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. In: *ICML*, pp. 179–186 (1997)
8. Estabrooks, A., Japkowicz, N.: A mixture-of-experts framework for learning from imbalanced data sets. In: Hoffmann, F., Hand, D.J., Adams, N., Fisher, D., Guimaraes, G. (eds.) *IDA 2001*. LNCS, vol. 2189, pp. 34–43. Springer, Heidelberg (2001). [https://doi.org/10.1007/3-540-44816-0\\_4](https://doi.org/10.1007/3-540-44816-0_4)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
10. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: *International Conference on Intelligent Computing*, pp. 878–887. Springer, Heidelberg, August 2005
11. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 475–482. Springer, Heidelberg, April 2009
12. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. In: *ICML*, vol. 97, pp. 179–186, July 1997
13. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. *Comput. Intell.* **20**(1), 18–36 (2004)
14. Wu, G., Chang, E.Y.: KBA: kernel boundary alignment considering imbalanced data distribution. *IEEE Trans. Knowl. Data Eng.* **17**(6), 786–795 (2005)
15. Ohsaki, M., Wang, P., Matsuda, K., Katagiri, S., Watanabe, H., Ralescu, A.: Confusion-matrix-based kernel logistic regression for imbalanced data classification. *IEEE Trans. Knowl. Data Eng.* **29**(9), 1806–1819 (2017)

16. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Networks Learn. Syst.* **29**(8), 3573–3587 (2017)
17. Dong, Q., Gong, S., Zhu, X.: Imbalanced deep learning by minority class incremental rectification. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(6), 1367–1381 (2018)
18. Breiman, L.: Bagging predictors. *Mach. Learn.* **24**(2), 123–140 (1996)
19. Ren, Y., Zhang, L., Suganthan, P.N.: Ensemble classification and regression-recent developments, applications and future directions. *IEEE Comput. Intell. Mag.* **11**(1), 41–53 (2016)
20. Martínez-Muñoz, G., Suárez, A.: Using boosting to prune bagging ensembles. *Pattern Recogn. Lett.* **28**(1), 156–165 (2007)
21. Li, Z.X., Zhao, L.D.: A SVM classifier for imbalanced datasets based on SMOTEBoost. *Syst. Eng.* **26**(5), 116–119 (2008)
22. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **40**(1), 185–197 (2009)
23. Guo, H., Viktor, H.L.: Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach. *ACM SIGKDD Explor. Newsl.* **6**(1), 30–39 (2004)
24. Hakim, L., Sartono, B., Saefuddin, A.: Bagging based ensemble classification method on imbalance datasets. *Int. J. Comput. Sci. Netw.* **6**, 7 (2017)
25. Yongqing, Z., Min, Z., Danling, Z., Gang, M., Daichuan, M.: Improved SMOTEBagging and its application in imbalanced data classification. In: *IEEE Conference Anthology*, pp. 1–5. IEEE, January 2013
26. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **42**(4), 463–484 (2011)
27. Rekha, G., Tyagi, A.K., Krishna Reddy, V.: A wide scale classification of class imbalance problem and its solutions: a systematic literature review. *J. Comput. Sci.* **15**, 886–929 (2019)
28. Rekha, G., Tyagi, A.K., Krishna Reddy, V.: Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method. *Int. J. Hybrid Intell. Syst.* **15**, 67–76 (2019)