KSMOTEEN: A Cluster Based Hybrid Sampling Model for Imbalance Class Data

Poonam Dhamal and Shashi Mehrotra

Abstract Classification accuracy for imbalance class data is a primary issue in machine learning. Most classification algorithms result in insignificant accuracy when used over class imbalance data. Class imbalance data exist in many sensitive domains such as medicine, finance, etc., where infrequent events such as rare disease diagnoses and fraud transactions are required to be identified. In these domains, correct classification is essential. The paper presents a hybrid sampling model called KSMOTEEN to address class imbalance data. The model uses a clustering approach, the K-means clustering algorithm, and combines the SMOTEEN technique. The experimental result shows, the KSMOTEEN outperforms some existing sampling methods, thus improving the performance of classifiers for class imbalance data.

Keywords Class imbalance · Classification · SMOTEEN · SMOTE

1 Introduction

Classification algorithms have many usages of prediction and data analysis in real-life applications. Most classifiers provide unusual accuracy when trained over class imbalance datasets. Samples with unequal distributed class in a dataset are called class imbalance dataset. When one class's sample size is considerably less/more than the other class for a given dataset, this is an imbalanced class problem [1–4]. Given a data D, samples $S = s_1, s_2, \ldots, s_n$, and attributes $A = a_1, a_2, \ldots, a_n$, one of the a_i among $a_1, a_2, \ldots, a_n$ is the class attribute, which is to be predicted. The minor class represents a lesser percentage of the samples in a dataset, whereas the majority class

P. Dhamal
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

S. Mehrotra (✉)
Department of Computer Science and Information Technology, Teerthanker Mahaveer University, Moradabad, India
e-mail: sethshashi11@gmail.com

A. E. Hassanien et al. (eds.), *International Conference on Innovative Computing and Communications*, Lecture Notes in Networks and Systems 731,
https://doi.org/10.1007/978-981-99-4071-4_51

represents more [5]. A model trained using an unbalanced dataset often performs poorly because instances of major class overwhelm and the minority class samples are ignored [6]. The dataset with undistributed class is associated with various applications, including fraud detection, text categorization, protein function prediction, medical diagnosis, signal processing, remote sensing, and image annotation. In such applications, samples of the minor class are more important and sensitive, and focus on the minor class is required [7, 8]. In these applications, incorrect classification of minority class data may result in a very high cost financially or in other ways [9]. For instance, incorrectly classifying a cancer patient may be a loss of human life as a cancer patient is misclassified as healthy and may not provide the medical care needed for the patient. In the mentioned cases, it is critical to appropriately classify the minority group, where the classifier tends to misclassify the minor class samples due to the small number of samples [10]. Sometimes minor class data are treated as an outlier [11].

However, classification algorithms do not perform well over class imbalance data due to the mentioned factor. Sampling-based approaches can be used with imbalanced datasets to get better classification results [12, 13]. The paper designs and presents a hybrid sampling model named KSMOTEEN to improve the poor classification accuracy of classification algorithms over the class imbalance. The proposed model integrates the K-means algorithm and the SMOTEEN sampling method. The K-means algorithm groups the objects based on similar features [14, 15]. It initially selects the random seeds as centroids, compares all the objects with the centroid based on their similarity, and the objects placed in the respective clusters [16].

2 Objective

To develop a model for improving classification results for class undistributed data. Our contributions to achieve the main goal are as follows:

1. Evaluate and compare classification techniques for class imbalance data before applying any sampling method and after using some existing sampling methods.
2. Design and develop a hybrid framework named KSMOTEEN for balancing the distribution of imbalanced class data.
3. Evaluate and compare KSMOTEEN against contemporary sampling methods.

The remaining portion is structured as following: in Sect. 3 related research are discussed, Sect. 4 describes the proposed model, and Sects. 5 and 6 describe experimental result and analysis.

3 Related Work

Many researchers have addressed the class imbalance issue and designed and proposed various solutions. This section discusses some research.

Wang et al. [17] suggested a new approach for merging the locally linear embedding algorithm (LLE) and the traditional SMOTE algorithm. Using an LLE mapping technique, they mapped the synthetic data points back to the original input space. Their approach outperforms classic SMOTE in terms of performance.

Das et al. [18] presented a new algorithm that combines reverse-neighbour-nearest neighbour (R-NN) and synthetic minority oversampling (SMOTE). R-SMOTE is used to extract significant data points from the minority class and synthesize new data points from the reverse nearest neighbours. Comparative analysis is done for the proposed algorithm and four standard oversampling methods. The empirical analysis shows that R-SMOTE produced better results than existing oversampling methods used for the experiment.

Lee et al. [19] used a different SMOTE method that merged the SMOTE algorithm with fuzzy logic. A fuzzy C-means algorithm can be used to identify membership degrees quickly. The suggested technique is evaluated using several benchmark datasets and exhibits promising results paired with support vector machine classifiers.

Tallo and Musdholifah [20] presented the SMOTE-simple genetic algorithm (SMOTE-SGA) for creating unequal amounts of synthetic instances. They applied a genetic algorithm at SMOTE, and classification results improved. Md Islam et al. [21], presented the SMOTE for the prediction of the success of bank telemarketing. SMOTE technique used to balance the dataset and then analyzed it using the Naive Bayes algorithm. It will help to find the best strategies for the improvement of the next marketing campaign.

Bajer et al. [22] compared various oversampling techniques over various real-life data. Also, it explores different interpretations of the algorithm in an attempt to show their behaviour.

Li [23] proposed the random-SMOTE (R-S) method for increasing the number of samples in the little class sample space randomly. As a result, the chances of improving minor class samples in data mining tasks to almost equal to those of the major class. Using a data mining integration process, they could balance five UCI imbalanced datasets. These datasets are classified using the logistic algorithm. It is observed that integrating R-S and logistic algorithm improves classifier performance significantly.

Rustogi and Prasad [24] proposed a hybrid method of classifying imbalanced binary data using synthetic minorities oversampling and extreme learning machines. They used five standard imbalance datasets for the performance evaluation of the model.

Han et al. [25] presented a new minority oversampling method. The borderline-SMOTE1 and Borderline SMOTE2 oversamples a small number of items towards the borderline.

Liu et al. [26] proposed a model called PUDL using only positive and unlabelled learning with dictionary learning. The model worked in two phases. First, they extracted negative samples from the unlabelled data to generate a negative class. The second phase designed a model Ranking support vector machine (RankSVM)-based to incorporate positive class samples. Patel and Mehta [27] reviewed modified K-means to increase the efficiency of the k-means clustering algorithm for preprocessing, cluster analysis, and normalization approaches. Three normalization techniques with outlier removal show the best and most effective results for Mk-means, performance analysis of computed MSE for Mk-means and Mk-means with three normalization techniques with outlier removal shows the best and most effective result for Mk-means, which generates minimum MSE and improves the efficiency and quality of result generated by this algorithm.

Chawla [28] proposed an approach based on a mix of the SMOTE algorithm and the boosting technique for learning from imbalanced datasets. SMOTE Boost generates synthetic samples from the rare or minority class, altering the updating weights and adjusting for skewed distributions in the process.

Gök et al. [29] proposed a model that works in two stages: in the first, no preprocessing was used, while in the second, preprocessing was stressed for improved prediction outcomes. The adjusted random forest algorithm and multiple preprocessing approaches reached 0.98 accuracies at the end of the investigation.

Nishant et al. [30] designed a model name HOUSEN to improve classification accuracy. The author used AdaBoost algorithms, random forest, and gradient, support vector machine for the experiment. The model shows a promising result.

4 Proposed Method

The KSMOTHEENN integrates the k-means clustering algorithm and the SMOTEENN sampling method. We selected the particular clustering algorithm, as from the literature survey, we analyzed that the K-means algorithm needs a number of clusters, and correct cluster number improves result accuracy. In our dataset, a number of clusters are known, i.e. two. In the second step, SMOTEENN is applied over minor class samples. The result of the classification algorithms; SVM, KNN, LR, AND NB are analyzed before the execution of any sampling method and after the execution of KSMOTEEN and some state-of-the-art undersampling, oversampling, sampling, and hybrid sampling models.

Figure 1 presents the work process diagram of our proposed model.

The proposed model KSMOTEEN first executes the K-means algorithm to group the samples as per the class and then the SMOTEEN sampling method is applied. In the first step, execute the K-means model; in the second step, execute the SMOTEEN. Algorithm1 presents the pseudocode of our proposed model; KSMOTEEN.

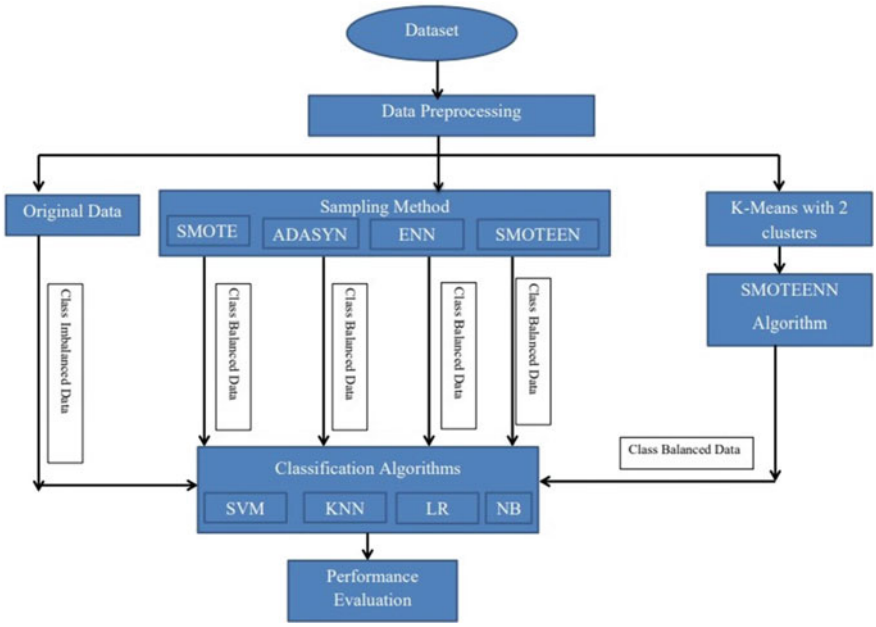

Fig. 1 Work flow process diagram

Algorithm 1 : KSMOTEEN Psudocode

Input : *Class imbalanced data*
Output: Class balanced data
Data cleaning and preprocess Initialize keyfeature list = []
ApplyK–Means clustering to the entire inputs place
Initialise the number of nearest numbers(k)
while end of majority sample **do**
  while kneighbors **do**
    Compute a line between the majority/minority data points and any of its neighbors and place a synthetic point
  end while end
while

5 Experiments

First, we preprocess the data and classification algorithms, SVM, KNN, LR, and NB, are executed over the data without applying any sampling method to see the impact of undistributed class data over classification results. We executed the mentioned classification algorithms over each class balance data obtained from mentioned sampling methods and the KSMOTEEN. Finally, we present comparative results.

5.1 Data Description

For the experiments, we used the UCI repository dataset. EEG data from students, and each student watched ten videos. As a result, it reduces the 12,000+ rows to just 100 data points. Each data point has more than 120 rows and is sampled every 0.5 s. For signals with a greater frequency, display the mean value every 0.5 s.

5.2 Evaluation Matrices

For the evaluation of the proposed model, we used the following measure [31, 32]:

Accuracy is a statistical measure that requires true positives and true negatives to estimate the model.

True positive (TP): No. of correctly classifying instances.

True negative (TN): No. of samples are identified as negative values correctly.

False positive (FP): No. of samples are wrongly predicted as positive.

False negative (FN): No. of samples are predicted incorrectly as negative.

Accuracy is defined mathematically as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{FP} + \text{TN}) \tag{1}$$

The recall is the percentage of instances of the class correctly identified.

$$\text{recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{2}$$

F1-score is the harmonic mean of precision and recall.

$$\text{F1 - score} = ((\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})) * 2 \tag{3}$$

Precision

$$\text{Precision} = \text{True Positives}/(\text{True Positives} + \text{False Positives}) \tag{4}$$

6 Result Analysis

We experiment in two phases. Phase I tests classification algorithms before applying any sampling method over imbalanced data and after using some existing mentioned sampling methods. Phase II demonstrates the results of the proposed model KSMOTEEN to evaluate the designed model's performance. The study used the following four classifier models: SVM, KNN, LR, and NB for the experiments.

From Table 1, it can be noticed that the accuracy of all the classification algorithms used for the experiment shows better results, except SVM. The SVM obtained better accuracy only in ADASYN, SMOTEN and K-means SMOTE. The KNN obtained better accuracy in the case of each sampling method. By analyzing Table 2, it can be observed the edited nearest neighbours.

Table 3 demonstrates that the accuracy of all the models improves after applying SMOTEEN and SMOTETomek, which are a combination of undersampling and oversampling methods.

Table 1 Accuracy of classification techniques before and after applying oversampling techniques

Sampling model	Classification techniques			
	SVM	KNN	LR	NB
Before sampling	58.69	54.97	58.97	54.48
Random oversampler	58.31	55.9	60.26	56.15
SMOTE	58.4	55.81	60.5	56.15
ADASYN	60.8	56.05	54.55	53.42
SMOTEN	59.16	56.36	50.79	56.42
Borderline SMOTE	57.79	55.54	59.98	56.05
K-means SMOTE	59.13	55.87	52.98	56.3
SVMSMOTE	57.88	56.12	60.01	56.03

Table 2 Accuracy of classification techniques before and after applying undersampling techniques achieved the best accuracy for all the classification algorithms

Sampling model	Classification techniques			
	SVM	KNN	LR	NB
Before sampling	58.69	54.97	58.97	55.57
Random undersampler	58.45	57.39	59.0	53.71
Cluster centroid	57.75	56.24	59.0	54.29
Condensed nearest-neighbour	65.62	56.79	65.11	65.03
Edited nearest neighbours	83.45	85.55	80.85	80.22
Neighbourhood cleaning rule	75.28	77.15	72.65	74.39
TomekLinks	61.65	61.27	61.34	59.77

At the same time, random undersampler did not perform with better accuracy for SVM and NB

Table 3 Accuracy of classification, accuracy before and after applying undersampling + oversampling techniques

Sampling model	Classification techniques SVM	KNN	LR	NB
Before sampling	58.69	54.97	58.97	55.57
SMOTEEN	88.28	80.21	77.93	64.99
SMOTETomek	61.23	60.23	61.01	55.27

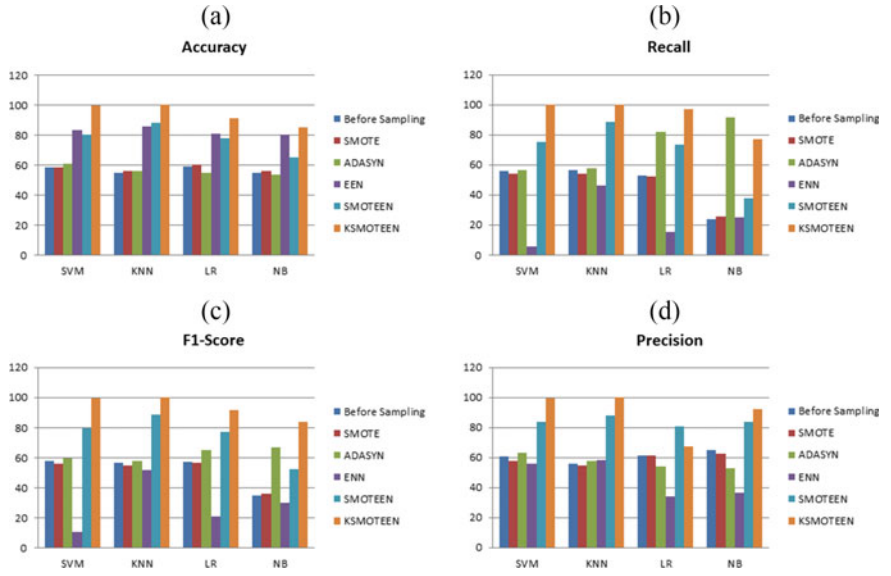

Fig. 2 Performance evaluation results of the classification models before and after applying the KSMOTEEN. Analyzing Fig. 2a–d, it is observed that all four classification algorithms' performance improved after the KSMOTEEN model's execution. However, the decision tree classification algorithm shows minor performance improvement after executing the KSMOTEEN model

Figure 2a–d demonstrates the experiments' accuracy, recall, f1-score, and precision before and after applying sampling methods and the proposed model, KSMOTEEN

7 Conclusion

In recent years, classification techniques have become increasingly popular for data analysis and prediction. Class imbalance is one of the primary issues for classifiers, due to which the performance of the classifier gets degraded. This paper presents a hybrid sampling model by integrating the K-means algorithm and SMOTEEN. The K-means technique is employed as an initial step to construct clusters. Further,

centroids are used by the KSMOTEEN. The KSMOTEEN model demonstrates promising results in improving the performance of classifiers. So, there is the scope that these techniques can be applied to any imbalanced dataset for accurate prediction purposes. Our future plan is to work for multi-class problems. Here, we have worked on data level approaches such as oversampling and undersampling. In the future, we plan to use algorithm level approaches.

References

1. Wasikowski M, Chen X-W (2009) Combating the small sample class imbalance problem using feature selection. IEEE Trans Knowl Data Eng 22(10):1388–1400
2. Dong Q, Gong S, Zhu X (2018) Imbalanced deep learning by minority class incremental rectification. IEEE Trans Pattern Anal Mach Intell 41(6):1367–1381
3. Mathew J et al (2017) Classification of imbalanced data by oversampling in kernel space of support vector machines. IEEE Trans Neural Netw Learn Syst 29(9):4065–4076
4. Bader-El-Den M, Teitei E, Perry T (2018) Biased random forest for dealing with the class imbalance problem. IEEE Trans Neural Netw Learn Syst 30(7):2163–2172
5. Beyan C, Fisher R (2015) Classifying imbalanced data sets using similarity based hierarchical decomposition. Pattern Recogn 48(5):1653–1672
6. López V et al (2013) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141
7. Hirsch V, Reimann P, Mitschang B (2020) Exploiting domain knowledge to address multi-class imbalance and a heterogeneous feature space in classification tasks for manufacturing data. Proc VLDB Endowment 13(12):3258–3271
8. Batista GEAPA, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newslett 6(1):20–29
9. Haixiang G et al (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239
10. Yong Y (2012) The research of imbalanced data set of sample sampling method based on K-means cluster and genetic algorithm. Energy Procedia 17:164–170
11. Siers MJ, Islam MZ (2020) Class imbalance and cost-sensitive decision trees: a unified survey based on a core similarity. ACM Trans Knowl Discovery Data (TKDD) 15(1):1–31
12. Tomek I (1976) Two modifications of CNN. IEEE Trans Syst, Man Cybern 6:769–772
13. Li Z, Kamnitsas K, Glocker B (2020) Analyzing overfitting under class imbalance in neural networks for image segmentation. IEEE Trans Med Imaging 40(3):1065–1077
14. Mehrotra S, Kohli S, Sharan A (2019) An intelligent clustering approach for improving search result of a website. Int J Adv Intell Paradigms 12(3–4):295–304
15. Mehrotra S, Kohli S (2017) Data clustering and various clustering approaches. In: Intelligent multidimensional data clustering and analysis. IGI Global, pp 90–108
16. Mehrotra S, Kohli S, Sharan A (2018) To identify the usage of clustering techniques for improving search result of a website. Int J Data Min, Model Manag 10(3):229–249
17. Wang J, Xu M, Wang H, Zhang J (2006) Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In: 2006 8th International conference on signal processing, vol 3. IEEE
18. Das R et al (2020) An oversampling technique by integrating reverse nearest neighbor in SMOTE: reverse-SMOTE. In: 2020 International conference on smart electronics and communication (ICOSEC). IEEE
19. Lee H et al (2017) Synthetic minority over-sampling technique based on fuzzy c-means clustering for imbalanced data. In: 2017 International conference on fuzzy theory and its applications (iFUZZY). IEEE

20. Tallo TE, Musdholifah A (2018) The implementation of genetic algorithm in smote (synthetic minority oversampling technique) for handling imbalanced dataset problem. In: 2018 4th international conference on science and technology (ICST). IEEE
21. Islam MS, Arifuzzaman M, Islam MS (2019) SMOTE approach for predicting the success of bank telemarketing. In: 2019 4th Technology innovation management and engineering science international conference (TIMES-iCON). IEEE
22. Bajer D et al (2019) Performance analysis of SMOTE-based oversampling techniques when dealing with data imbalance. In: 2019 International conference on systems, signals and image processing (IWSSIP). IEEE
23. Li J, Li H, Yu J-L (2011) Application of random-SMOTE on imbalanced data mining. In: 2011 Fourth international conference on business intelligence and financial engineering. IEEE
24. Rustogi R, Prasad A (2019) Swift imbalance data classification using SMOTE and extreme learning machine. In: 2019 International conference on computational intelligence in data science (ICCIDS). IEEE
25. Han H, Wang W-Y, Mao B-H (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. Springer, Berlin, Heidelberg
26. Liu B, Liu Z, Xiao Y (2021) A new dictionary-based positive and unlabeled learning method. Appl Intell 51(12):8850–8864
27. Patel VR, Mehta RG (2011) Impact of outlier removal and normalization approach in modified k-means clustering algorithm. Int J Comput Sci Issues (IJCSI) 8(5):331
28. Chawla NV et al (2003) SMOTEBoost: improving prediction of the minority class in boosting. In: European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg
29. Gök EC, Olgun MO (2021) SMOTE-NC and gradient boosting impu- tation based random forest classifier for predicting severity level of covid-19 patients with blood samples. Neural Comput Appl 33(22):15693–15707
30. Nishant PS et al (2021) HOUSEN: hybrid over–undersampling and ensemble ap- proach for imbalance classification. In: Inventive systems and control. Springer, Singapore, pp 93–108
31. Wegier W, Koziarski M, Wozniak M (2022) Multicriteria classifier ensemble learning for imbalanced data. IEEE Access 10:16807–16818
32. Brzezinski D et al (2019) On the dynamics of classification measures for imbalanced and streaming data. IEEE Trans Neural Netw Learn Syst 31(8):2868–2878