(ISSBM) Improved Synthetic Sampling based on Model for Imbalance Data

Ragini Gour
M.Tech Scholar CSE
SATI, Vidisha, M.P

Ramratan Ahirwal
Assistant Professor
Deptt of CSE
SATI, Vidisha, M.P

ABSTRACT
In the data mining research domain imbalanced data is characterized by the rigorous variation in scrutiny frequency between classes and has expected a lot of consideration. The forecast performances usually depreciate as classifiers learn from data imbalanced, as most of classifiers presume the class division is balanced or the costs for different types of classification errors are the same. Although several methods have been analyzed to deal with imbalance problems, it is still difficult to oversimplify those methods to achieve stable improvement in most cases. In this study, we propose a novel framework called Improved Synthetic Sampling Based on Model (ISSBM) to deal with imbalance problems, in which we integrate improved modeling and sampling techniques to generate synthetic data. The key inspiration behind the proposed method is to use deterioration models to capture the relationship between features and to consider data multiplicity in the process of data generation. We conduct experiments on many datasets and compare the proposed method with 5 methods. The experimental results indicate that the proposed method is not only qualified or comparative but also very stable. We also provide detailed analysis of the proposed method to empirically demonstrate why it could generate good data samples.

Keywords
Imbalance data, random over sampling, random under sampling, synthetic minority over sampling technique

1. INTRODUCTION
We can say machine learning is concept of how to make computer more powerful. The last decade has witnessed the success of machine learning outstanding to the explosion of data and advancements in computing power. Machine learning(ML) has been successfully functional on many application domains, including, the medical field domain [6], finance field domain [7], and manufacturing field domain [8] but not limited. As machine learning classifiers learn from imbalanced data, their prediction performances often deteriorate significantly [2]. This is because most machine learning algorithms presume that the underlying class circulation is balanced [9] or the costs for different errors of classification are equal [2]. Therefore, the data imbalanced problem would bias the classification assessment toward the majority class. As we are always fascinated about the minority class (e.g., the positive case for medical verdict and fault event for manufacturing), the aforesaid problems really cause a significantly impact in practice.

Although many methods handling imbalanced data have been proposed from many years, it is difficult to describe and generalize them to achieve stable improvement in most cases. The sampling technique and method is probably one of the most widely used methods in dealing with imbalanced data, as it is easy to implement. One drawback of the sampling method is that data samples are increased or decreased without considering the underlying data distribution. Imbalance data Over-sampling technique may possibly result in an overfitting problem, while under-sampling may reject representative samples. Furthermore, many synthetic sampling approaches produce synthetic samples based on k-nearest neighbors, which may be biased by the samples in the minority class.

In this work, we propose and recommended a improved synthetic sampling based on model (ISSBM) method, which is a new framework that oversamples the minority class instances from a new aspect. The proposed model belongs to the over-sampling technique of imbalance data, and the goal is to generate synthetic samples that could capture the relationship between the features of training samples that are in the minority class, while keeping the variability of the data samples. Compared with previous methods, the proposed method generates synthetic samples based on several ideas. First, the proposed method uses the modeling technique to capture trends or regression lines of the features for the training samples in the minority class.

Although second it produce provisional data samples by sampling available feature values. Finally, it convert and transforms temporary data samples into synthetic data via the constructed model. In the experiments, we analyze and compare the proposed method with several alternatives on few datasets, and the experimental results indicate that the proposed method is not only effective and comparative, but also stable and better than other approach. We also provide detailed investigations and visualizations of the proposed method of imbalance data to empirically demonstrate why it could generate good data samples. The contributions of this work are listed as follows. First, we focus on imbalanced data problems and propose a model-based improved synthetic sampling method. Second, we design several experiments to assess the proposed method. We conduct experiments on thirteen datasets and compare the results of the proposed method with those of ten and more competitive technique and methods. The experimental results shows that the suggested and proposed method is comparative and outperforms other alternatives in most cases. Finally, we provide detailed investigations and use a visualization method and technique to empirically show that the proposed method performs well when compared with all the other alternatives methods. Moreover, the proposed method is a data-level algorithm, which is easy and stretchy to extend. We combine the proposed method with the boosting technique to devise a method called ISSBMBoost, and compare the performance of the combination with two state-of-the-art ensemble-based methods with regards to the imbalance problem. The

experimental results indicate that the ISSBMBoost works well and stably on the thirteen datasets. The rest of this paper is organized as follows. Section 2 presents correlated surveys about the imbalanced data and the methods. Section 3 then introduces the proposed model. Next, Section 4 summarizes the experimental and investigating settings and results.Section 5 conclusion.

Any dataset with irregular distribution and sharing between its minority and majority classes can be considered to have class imbalance, and in the real-world applications, the severity of class imbalance can vary from minor to severe (high or extreme). A dataset can be considered imbalanced if the classes, e.g., non-fraud cases and fraud cases, are not equally represented. The majority class makes up most of the dataset, whereas the minority class, with limited dataset representation, is often considered the class of interest. With the real-world datasets, class imbalance should be expected. If the degree of class imbalance for the majority class is tremendous, then a classifier may yield high overall prediction precision since the model is likely predicting most instances as belonging to the majority class. Such a model is not practically useful, since it is often the prediction performance of the class of interest (i.e., minority class) that is more important for the domain experts [1]. He and Garcia [2] suggest that a popular viewpoint held by academic researchers defines imbalanced data as data with a high-class imbalance between its two classes, stating that high-class imbalance is reflected when the majority-to-minority class ratio ranges from 100:1 to 10,000:1. While this range of class imbalance may be observed in big data, it is not a strict definition of high-class imbalance. From the viewpoint of effective problem-solving, any class imbalance (e.g., 50:1) level that makes modeling and prediction of the minority class a complex and challenging task can be considered high-class imbalance by the domain experts [3]. It should be noted that we focus our survey investigation of published works on class imbalance in big data in the context of binary classification problems, since typically non-binary (i.e., multi-class) classification problems can be represented using a sequence of multiple binary classification tasks.

The scope of our study is investigating works conducted within the past 8 years (i.e., 2010–2018) that focus on the problem junction of big data and class imbalance, and the consequent solutions developed by researchers. Moreover, in the interest of our focus on big data only, we only consider relevant works that analyze (class imbalance in big data) at least one dataset consisting of 100,000 instances or higher. In addition to analyzing the surveyed papers, we also provide our own insights into likely gaps in current research in the area and discuss avenues for future work for the community. To the best of our knowledge, we have included all published articles that fall within our survey study's scope. We believe such a large-scale survey of works addressing, and developing solutions for, high-class imbalance problems in big data is unique in the data mining and machine learning domain.

In related literature, the strategies for tackling class imbalance problems are similar for both traditional data and big data. Ali et al. [8] categorize approaches for addressing class imbalance into those conducted at the Data-Level or at the Algorithm-Level, where both categories include approaches used for both traditional data (see Fig 1), i.e., data sampling (over-sampling and under-sampling), feature selection, cost-sensitive methods, and hybrid/ensemble techniques. Data-Level methods include feature selection approaches and data sampling, while Algorithm-Level approche and methods includes hybrid/ensemble and cost-sensitive approaches. This categorization is further defined in the next section.

Based on our study investigation we observed some interesting trends/results of the surveyed works, and some important keys findings are summarized next. Among the Data-Level methods, pragmatic results of relevant works generally suggest that Random Over-Sampling (ROS) yields better classification performance than Random Under-Sampling or the Synthetic Minority Over-Sampling Technique (SMOTE). Moreover, with the MapReduce environment for big data analysis with data sampling, the process of determining the preferred balance between the data over-sampling percentage and classification performance is an empirical parameter/process, instead of a formulaic solution. At the Algorithm-Level, there are a variety of methods that seemingly provide good classification performance for big data with high-class imbalance. For the cost-sensitive techniques of imbalance data, our discussion includes a fuzzy rule-based classification approach [9, 10] and an online learner scheme [11, 12]. For the hybrid/ensemble techniques, our discussion includes a Bayesian Optimization Algorithm that maximizes Matthew's Correlation Coefficient by learning optimal weights for the positive and negative classes [13], and an approach that combin es Random Over-Sampling(ROS) approachand Support Vector Machines (SVMs) approach [14].

One of the most important primary problems we encountered throughout our detailed surveyed works was that the MapReduce big data framework was experiential to be very sensitive to the high-class imbalance [15], primarily due to the undesirable effects of creating several partitions within the already very small minority class space. Hence, we suggest a superior focus on a more flexible computational atmosphere for big data analysis, such as Apache Spark [16], for addressing the high-class imbalance problem. Another key issue plaguing big data is small disjuncts (described later in the paper) of data points within the overall dataset or within each of the two classes, and based on our survey, we note that this issue has not been given enough focus in the context of high-class imbalance in big data. In addition, given the problem's fairly deprived maturity in developed effective solutions, considerably more research and empirical investigation still remain to be conducted. Many studies we investigated in this paper generally lacked sufficient deepness in the scope of their empirical investigation of the high-class imbalance problem in big data. This finding makes it difficult to conclude whether one approach is more efficient and effective than another approach.

The main and important purpose of this article is organized as follows. In "All Methods and technique addressing data imbalance in traditional data" section, we provide an significant overview of methods and strategies for handling traditional data with the class imbalance problem. While the primary focus of this paper is on high-class imbalance in big data, we present "Methods addressing class imbalance in traditional data" section to provide the reader with a more complete picture of existing approaches for class imbalance, since similar methods are generally used for both traditional data and big data. In "Methods addressing class imbalance in big data" section, we discuss the Data-Level methods and Algorithm-Level techniques for handling big data defined by high degrees of class imbalance.In "Conclusion" section, we conclude with the main points of our paper and suggest some directions for future work.

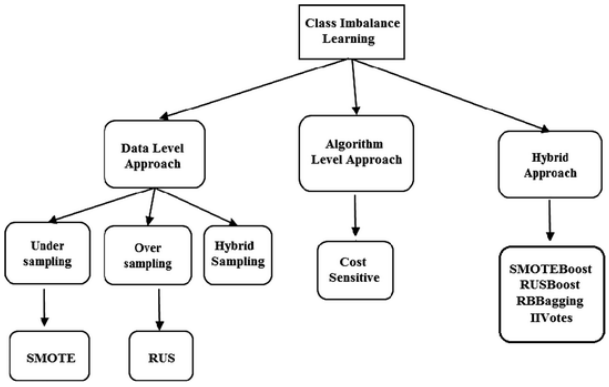

Fig 1: Class Imbalance techniques

2. RELATED WORKS

Imbalance data problems are unsafe to many kinds of classifiers. Sun et al. (2009) [10] have studied the difficulties of learning from data imbalanced for several machine learning algorithms, including support vector machines (SVM) [11] decision tree, artificial neural network (AI). Consequently, plentiful researchers have dedicated time to designing methods for imbalanced data, and these methods could be categorized into two types, algorithm level and data-level methods [10]. We also conducts a literature survey of ensemble-based approaches for the imbalance data problem, as applying the ensemble learning technique to handle imbalance problems has become popular in recent years.

2.1 Data-level Approach

In the data-level approach of imbalance data adjusts the class distribution by the resampling original data and producing synthetic data to remedy imbalance problems. As this approach is always useful to data before constructing a classification model, it could be regarded as part of the pre-processing step. The main and important benefit of the data-level approach is that the methods utilized are self-sufficient of classifiers. The resampling approach for imbalanced data originated from the study of Kubat and Matwin [12]. They can be roughly categorized into under-sampling and over-sampling methods. The aim of over-sampling is to raise the number of samples in the minority class by resampling or generating synthetic data, while that of under-sampling is to decrease the number of samples in the majority class by removing samples from it. The original resampling methods are random over-sampling and random under-sampling [13]. Random over-sampling(ROS) choose minority data by random sampling, i.e., the probability of each sample is the same, with substitute and then adds the selected samples to the original dataset. Random under-sampling selects samples from majority set without replacement and then removes the selected data from the dataset.

2.2 Algorithm-level Approach

The approach at the algorithm-level avoids the imbalance data problem by directly modifying the classifier or using diverse mis-classified costs to improve the entire performance of the model. These methods mediate in the training stage, so most of them are not self-sufficient of classifiers. Compare with the data-level approach operates at the data level; therefore, a variety of classifiers could benefit from it. The essential purpose of cost-sensitive learning is to use a cost matrix to adjust the penalties for various errors. In most settings, we are fascinated in the minority class. Given that the minority class is the positive class and majority class is the negative class, the penalty or cost for false negatives (FN) is higher than that for false positives (FP), so that the learning algorithm can emphasize the importance of the minority. Wu and Chang [26] anticipated a method called kernel-boundary alignment (KBA), which redesigns the kernel trick according to the imbalanced data distribution. In this process one of the most popular kernel functions is the radial basis function (RBF), and the KBA is based on RBF as well. The kernel function for KBA involves not only the distance between all data points and the support vectors, but also the class distribution of the support vectors. Thus, the imbalanced situation is under consideration in the training process.

On the other hand based on kernel regression logistic, Ohsaki et al. [27] suggested a method called confusion-matrix-based kernel logistic regression (CM-KLOGR) for imbalanced data. The key idea is to include the weighted harmonic mean of various performance metrics from the confusion matrix in the loss function. The CM-KLOGR involves two steps. The first step is to pre-train a model with cross-entropy loss, which is the same as the original kernel regression logistic. Furthermore, second step is to retrain the model, initialized by the pre-training parameters, with their proposed loss function to simultaneously consider various performance metrics for imbalanced data, so that the minority class will not be unnoticed.

In the Algorithm-Level approach can be further divided into cost- hybrid/ensemble method and sensitive methods. The former works on the general principal of assigning more weight to an instance or learner in the event of a misclassification, e.g., a false negative prediction may be assigned a higher cost (i.e., weight) compared to a false positive prediction, given the latter is the class of interest. Ensemble methods can also be used as cost-sensitive methods, where the classification outcome is some combination of multiple classifiers built on the dataset; Bagging and Boosting are two common types of ensemble learners [24, 25]. Bagging minimizes the predictive variance by producing several training sets from the given dataset, with a classifier being generated for each training set and then their individual models combined for the final classification. Boosting also uses several training sets from the given dataset, and after iteratively assigning different weights to each classifier based on their misclassifications, a weighted approach combining the individual classifier's results yields the final classification. Hybrid methods are designed to mixture known problems arising from the data-sampling methods, feature selection methods, cost-sensitive methods, and basic learning algorithms such as Naive Bayes [26]. In some instances, sub-groups of Data-Level methods or Algorithm-Level methods may be combined into an overall approach to address the class imbalance problem. For example, the popular Random Forest (RF) classifier is a version of the original Random Decision Forest [27] algorithm, and is an ensemble learner which also implements Bagging. In contrast, the original Random Decision Forest is not considered an ensemble learner [28].

2.3 Ensemble Approach

Ensemble learning is a machine learning approach that trains a set of hypotheses and combines them to make a prediction. The most common methods in ensemble learning include boosting [33], bagging (bootstrap aggregating) [34], and stacking [35]. Several researchers have applied boosting to imbalance problems and devised novel algorithms that integrate either oversampling or under-sampling methods into ensemble learning framework. The boosting approach is similar to the bagging approach in terms of constructing multiple classifiers by sampling, but boosting tends to select

the samples that are mis-classified by previous classifiers in general.

3. IMPROVED SYNTHETIC SAMPLING BASED ON MODEL

This segment introduces the projected method, including the algorithm, notation and our motivation in this research work.

3.1 Concept and Motivation

The over-sampling approach and method require replication of minority samples class, primary to many overlapping shows in the feature space. Ideally, the importance of the minority data is superior, but the small disjunct problem is still rigorous, as the scope of the minority data is not distended. This condition forces the classifier to develop a large number of precise and narrow decision regions to classify each point appropriately, but it is difficult for those decisions to be comprehensive to unseen data [14]. Main purposes of SMOTE is to deal with the aforementioned problem. SMOTE achieved this by producing synthetic data to enlarge the extent of the minority data in the feature space, so that the generalization may be enhanced.

3.2 Notation

In this segment, the notations that we will be used in the following sections are introduced. Each data sample object $x^{(i)}$ is represented as a feature vector of length m, i.e., $x^{(i)}=[x_1,x_2.....x_m]$. x_{-j} is introduced to represent all the features of data sample x except feature j. The data compilation is represented by $D=x^{(1)},x^{(2)}.....x^{(n)}$, indicating that the number of data objects is n and $x^{(i)}$ belogs to R^m. To shorten the explanation, we center of attention on binary classification, but the projected work could be extended to multi-class problems. The focus of this study is imbalanced data; therefore, data could be separated into two classes, namely, the majority class and minority class. We use D^s to denote the minority class samples and use D^l to represent majority class data samples. Thus, the entire data set can be divided into two partitions, namely, $D=D^s U\ D^l$. We propose to generate temporary synthetic data by sampling features, so we further introduce the value set $V=\{v_1,v_2.....v_m\}$ for all the possible feature values of the minority class, so that for all $x^{(i)}$ belongs to D^s

3.3 Proposed Method

The goal of ISSBM is to produce synthetic data that preserve the characteristics of the original data. In this work, we presume there exists a relationship between features, and the relationship could be used to hold the characteristics of the features. Thus, we suggest and propose to build models to study the relationship between features and obtain synthetic data in terms of the learned feature models. ISSBM involves three steps: (a) We used training feature models,(b) We used sampling features to produce temporary sampling data, and (c) generating final synthetic data. This section introduces each step in detail and shows the proposed algorithm. The proposed method and approach is a framework, and the feature models could be replaced with different models. In the implementation, we use non-linear regression and linear models as the feature models.

3.4 Proposed Algorithm

Algorithm 1 shows that the inputs include the over-sampling rate Z, minority data , the feature value set V, and the number of iterations for repeating the generation process T. First, the temporary sampling data set R and the synthetic data set SD are initialized as empty matrices of size m shown in Line 4 and Line 5 of Algorithm 1. Next, we train m feature models for the m categorized features. Lines 6-8 show the steps for training feature models. The second step of the proposed algorithm is to produce temporary data samples by applying the random with replacement technique on the features as presented in Lines 9-13. The final step is to produce synthetic data samples as listed in Lines 14-19, in which SDj denotes the jth feature of the synthetic data sample and is obtained from the prediction of modelj with R as the input. The proposed algorithm is a data-level algorithm; therefore the outputs are the synthetic data samples. Once the generation process is completed, we use the final synthetic data as well as the original data to train the classifier.

Algorithm 1. ISSBM Algorithm

Input: D^s : minority data sample, Z: over-sampling rate, $V=[v_1,v_2.....v_m]$: possible feature values of minority classes, T: total iterations for repeating generation process

Output: SD: synthetic data sample

1 n^s = number of minority class samples
2 m = number of categorized features
3 n^{syn} = n^s * Z (synthetic data size**)**
4 R = (temporary data with **n^{syn}** rows and m columns)
5 SD= (synthetic data with m columns and **n^{syn}** rows)
6 for j = 1 to m do
7 train modelj with xj as label and xj as features
8 end
9 for i = 1 to **n^{syn}** do
10 for j = 1 to m do
11 Rij = randomly sample a value from vj with replacement
12 end
13 end
14 for t = 1 to T do
15 for j = 1 to m do
16 Sj = predict feature j by modelj and R
17 end
18 R = S (update the temporary dataset for predicting by predicted dataset S)
19 end
20 return S

3.5 Discussion

The proposed approach is a strategy and framework that contain three steps. The first and final steps are related to regression models, while the aim of the second step is to arbitrarily produce temporary synthetic samples. We momentarily provide the motivation for using these two components to discuss and studied their effectiveness in the proposed method. In this work, we propose a data-level algorithm that utilizes the over-sampling technique to produce synthetic samples. Methods that produce synthetic samples to balance class distribution usually build specific assumptions during the generating process. For example, SMOTE assumes that the synthetic samples are located in the line between two minority samples, so it is expected that SMOTE considers local information to produce synthetic samples. Although, some researchers presume that the minority samples should be produced on a new categorized feature space that could hold structure information [24], [25]. To produce synthetic samples

with good and effective quality, we presume there exists a relationship between categorized features that could be characterized by regression models. Thus, we have to use simple models to capture the relationship, as the number of minority samples limited as always in an imbalance data problem. We endeavor to learn the trends or regression lines of the features from minority samples, so that the regression lines can facilitate the production of final synthetic samples when given provisional synthetic samples. It is worth mentioning that randomness is an important component in sampling-based methods. In the second step, the provisional sampling data are generated randomly, so as to boost the diversity of the synthetic data. According to

Table 1: Summary Description of Dataset

Name	Data size	# of Features	Imbalance ratio
Pima Indian Diabetes	768	8	268:500(34.9%)
Haberman's Survival	306	3	81:225(26.5%)
Satimage	6435	36	626:5809(9.7%)
E.coli	336	7	35:301(10.4%)
Shuttle	43500	9	37:43463(0.09%)
Bank	45211	11	5289:39922(11.7%)

our survey, Guo and Viktor have used similar approaches to generate synthetic samples to balance class distributions as described in Section 2.2 of [37]. The key idea behind data generation in our work and [37] is similar, namely, to generate diverse synthetic samples. The data generation step in the proposed method is similar to the rule of nominal attribute proposed by Guo and Viktor [37]. We do not make an assumption about the distribution underlying the original training attributes; instead, we use models to adjust the attribute values in order to emulate the real feature relationships. Guo and Viktor applied this technique to deal with an imbalanced data problem [37], and improve classification performance by using boosting with data generation. We conduct experiments to compare the proposed framework with several alternatives, and investigate the effectiveness of the steps involved in the proposed framework.

4. EXPERIMENT AND DISCUSSION

4.1 Datasets

We conducted experiments on many datasets to evaluate the performance of the proposed method and other competitive methods. The summaries of the datasets are listed in Table 1, including the data size, number of features, and imbalance ratio. All of the data used in the experiments are publicly available datasets.

4.2 Evolution Matric

Accuracy is probably one of the most commonly used performance metrics for classification tasks. However, the limitation of accuracy as the performance measure on an imbalanced dataset was quickly established, and receiver operating characteristic (ROC) curves soon emerged as a popular choice in which the x-axis is the false positive rate (FPR) and the y-axis is the true positive rate (TPR). To compare classifiers, one may want to reduce ROC performance to a single scalar value representing expected performance. The area under the curve (AUC) is an alternative method for evaluating classifier performance, explaining why this work uses AUC as the evaluation metric

4.3 Experimental Setting

As conduct experiments, we compared the proposed method with various other competitive methods, including random oversampling method, random under-sampling method, SMOTE, borderlineSmote1, safe-level-SMOTE, ADASYN, cluster-based over-sampling method, and the under-sampling method based on clustering. All these methods are classical and frequently used in handle imbalanced data, explaining why they were chosen in the experiments. Sampling methods fit in to the stochastic process, so two confounding factors may lead to unsteady experimental results: (a) One is different random seeds for splitting training and testing data, and (b) Second different random seeds for performing sampling method. Consequently, we proposed a more precise design in our experiments as illustrated in Fig. 4, in which we used 8 different seeds to split data into testing data and training, and performed the sampling methods for each splitting data with another 8 different seeds. In each data splitting, we could collect 8 performance results for each method, and we used the average performance as the result for this splitting. Then, we averaged the performance from 8 splittings to get the final performance result for each method; in other words, the final performance result was based on 80 experimental results. The projected method and the viable methods are data-level algorithms, and the experiments focused on a classification task, so the experiments needed a classifier to perform the classification work. We used logistic regression as the classifier, as it is a normally used classification algorithm. Although, the proposed method is a framework, and we can use diverse modeling methods as our feature models. The aim of the feature model is to discover the relationship between features, so any model could be used in the proposed method. This work focuses on numeric features, so the feature models are regression models. In imbalance data most over-sampling methods also focus on numeric features to produce data samples. Following the setting of the classification algorithm, we chosen a linear algorithm and two non-linear algorithms as the feature models to assess the performance of the proposed method, in which the linear

4.4 Experimental Result

The experimental results are shown in Table 6 and the rankings of the variants are listed in Table 7. The experimental results specify that ISSBM generally outperforms "No modeling". The difference between these two methods is the modeling process, and the results specify that ISSBM could gain from the modeling step. Next, the evaluation for the second step requires a comparison of ISSBM and the variants of ISSBM. It is apparent that ISSBM outperforms these three alternatives, indicating that sampling for each feature is an effective step to improve performance. Finally, we examined the performance differences of ISSBM at different iteration numbers to verify the effect of the iterative step used in the final step of ISSBM. The average AUC on four datasets with different numbers of iterations. The experimental results indicate that the proposed method performs stably as the number of iterations increases, and could achieve better performance on the Vehicle dataset.

Table 2: Average AUC for Different Variations of ISSBM

Dataset	ISSBM	No modeling	ISSBM (Data Level Sampling)	ISSBM(Random Sampling)	ISSBM (No sampling)
Pima	0.8254	0.8248	0.8246	0.825	0.8247
Haberman	0.6873	0.6818	0.6873	0.6874	0.6874
Satimage	0.7645	0.7628	0.7632	0.7273	0.7635
Ecoli	0.9296	0.9262	0.9269	0.9214	0.9278
Shuttle	0.8902	0.8938	0.8687	0.8625	0.8767
Vehicle	0.9882	0.9863	0.9861	0.9872	0.9851

Table 3: Average AUC for Different Variations of ISSBM

Dataset	ISSBM	No modeling	ISSBM (Data Level Sampling)	ISSBM(Random Sampling)	ISSBM (No sampling)
Pima	1	2	4	3	4
Haberman	3	5	3	1	1
Satimage	1	4	3	5	2
Ecoli	1	4	3	5	2
Shuttle	2	1	4	5	3
Vehicle	1	3	4	2	4

5. CONCLUSION

As this research work going from many steps we can conclude that imbalance data problems have occur in a variety of application domains and established a considerable amount of focus recently. This work proposes a new concept called ISSBM to deal with imbalance problems. The proposed work combine sampling and modeling techniques to produce synthetic data, and the generating process involves three steps. We conducted experiments on many datasets and compare the proposed approach and method with ten competitive methods. The experimental results signify that the proposed method outperforms other alternatives approaches in most cases in terms of robustness and effectiveness. The proposed method is a framework, and many future study and research instructions are possible, such as applying other regression models and other classifiers to diverse application domains for further study and analysis. Although, the effect of various over-sampling rates requires to be explored to find the most suitable setting according to various conditions. Finally, although we conducted experiments on several datasets in this work, it is worthwhile to focus on a certain application domain to deeply explore the relationship between features to refine the feature model structure.

6. REFERENCES

[1] Bauder RA, Khoshgoftaar TM. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced Big Data. Health Inf Sci Syst. 2018.

[2] Triguero I, Rio S, Lopez V, Bacardit J, Benítez J, Herrera F. ROSEFW-RF: the winner algorithm for the ECBDL'14 big data competition: an extremely imbalanced big data bioinformatics problem. Knowl Based Syst. 2015.

[3] Van Hulse J. A study on the relationships of classifier performance metrics. In: 21st international conference on tools with artificial intelligence (ICTAI 2009). IEEE. 2019.

[4] Katal A, Wazid M, Goudar R. Big data: issues, challenges, tools, and good practices. In: Sixth international conference on contemporary computing. 2013.

[5] Herland M, Khoshgoftaar TM, Bauder RA. Big Data fraud detection using multiple medicare data sources. J Big Data. 2018;5:29

[6] Bauder RA, Khoshgoftaar TM. Medicare fraud detection using random forest with class imbalanced Big Data. In: 2018 IEEE international conference on information reuse and integration (IRI), IEEE. 2018. pp. 80–7.

[7] Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem: a review. Int J Adv Soft Comput Appl. 2015;7(3):176–204.

[8] Lopez V, Rio S, Benitez J, Herrera F. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced Big Data. Fuzzy Sets Syst. 2015;258:5–38.

[9] Wang D, Wu P, Zhao P, Hoi S. A framework of sparse online learning and its applications. Comput Sci. 2015.

[10] Zhang T. Sparse online learning via truncated gradient. J Mach Learn Res. 2018.;10:777–801.

[11] Maurya A. Bayesian optimization for predicting rare internal failures in manufacturing processes. In: IEEE international conference on Big Data. 2016.

[12] Galpert D, del Río S, Herrera F, Ancede-Gallardo E, Antunes A, Agüero-Chapin G. An effective Big Data supervised imbalanced classification approach for ortholog detection in related yeast species. BioMed Res Int. 2015;2015:748681.

[13] Tsai C, Lin W, Ke S. Big Data mining with parallel computing: a comparison of distributed and MapReduce

methodologies. J Syst Softw. 2016;122:83–92.

[14] Triguero I, Galar M, Merino D, Maillo J, Bustince H, Herrera F. Evolutionary undersampling for extremely imbalanced Big Data classification under Apache Spark. In: IEEE congress on evolutionary computation (CEC). 2016.

[15] Khoshgoftaar TM, Seiffert C, Van Hulse J, Napolitano A, Folleco A. Learning with limited minority class data. In: Sixth international conference on machine learning and applications (ICMLA 2007), IEEE. 2007. pp. 348–53.

[16] Malhotra R. A systematic review of machine learning techniques for software fault prediction. Appl Soft Comput. 2015;27:504–18.

[17] Wang H, Khoshgoftaar TM, Napolitano A. An empirical investigation on Wrapper-Based feature selection for predicting software quality. Int J Softw Eng Knowl Eng. 2015;25(1):93–114.

[18] Yin L, Ge Y, Xiao K, Wang X, Quan X. Feature selection for high-dimensional imbalanced data. Neurocomputing. 2013;105:3–11.

[19] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. Explor Newsletter. 2014;6(1):80–9.

[20] Seiffert C, Khoshgoftaar TM. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man Cybern Part A. 2010;40(1):185–97.

[21] Graczyk M, Lasota T, Trawinski B, Trawinski K. Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. In: Asian conference on intelligent information and database systems. 2010. pp. 340–50.

[22] Breiman L. Random forests. Mach Learn. 2015.;45(1):5–32.

[23] Ho T. Random decision forests. In: Proceedings of the third international conference on document analysis and recognition. 2016..

[24] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.

[25] Chawla N, Lazarevic A, Hall L, Bowyer K. SMOTEBoost: improving prediction of the minority class in boosting. In: 7th European conference on principles and practice of knowledge discovery in databases. 2013.

[26] Rodriguez D, Herraiz I, Harrison R, Dolado J, Riquelme J. Preliminary comparison of techniques for dealing with imbalance in software defect prediction. In: Proceedings of the 18th international conference on evaluation and assessment in software engineering. Article no. 43. 2014.

[27] Fernandez A, Rio S, Chawla N, Herrera F. An insight into imbalanced Big Data classification: outcomes and challenges. Complex Intell Syst. 2017;3:105–20.

[28] Cao P, Zhao D, Zaiane O. An optimized cost-sensitive SVM for imbalanced data learning. In: Pacific-Asia conference on knowledge discovery and data mining. 2013. pp. 280–92.

[29] Cao P, Zhao D, Zaiane O. A PSO-based cost-sensitive neural network for imbalanced data classification. In: Pacific-Asia conference on knowledge discovery and data mining. 2013. pp. 452–63.

[30] Li N, Tsang IW, Zhou Z-H. Efficient optimization of performance measures by classifier adaptation. IEEE Trans Pattern Anal Mach Intell. 2013;35(6):1370–82.

[31] López V, Fernandez A, Moreno-Torres J, Herrera F. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. Expert Syst Appl. 2012;39(7):6585–608.

[32] Kaminski B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. CEJOR. 2017;26(1):135–59.

[33] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets. In: European conference on machine learning. 2014.. pp. 39–50.

[34] Tang Y, Chawla N. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern. 2019.;39(1):281–8.

[35] Bekkar M, Alitouche T. Imbalanced data learning approaches review. Int J Data Mining Knowl Manag Process. 2013;3(4):15–33.

[36] 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 C Appl Rev. 2012;42(4):463–84.

[37] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach," ACM SIGKDD Explorations Newslett., vol. 6, no. 1, pp. 2014.