I.J. Intelligent Systems and Applications, **2018, 5, 71-81**
Published Online May 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2018.05.08

A Hybrid Approach for Class Imbalance Problem in Customer Churn Prediction: A Novel Extension to Under-sampling

Uma R. Salunkhe
Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune, 411041, India
E-mail: umasalunkhe@yahoo.com

Suresh N. Mali
Sinhgad Institute of Technology and Science, Pune, India
E-mail: snmali@rediffmail.com

Received: 12 May 2017; Accepted: 06 July 2017; Published: 08 May 2018

Abstract**—**Customer retention is becoming a key success factor for many business applications due to increasing market competition. Especially telecom companies are facing this challenge with a rapidly increasing number of service providers. Hence there is need to focus on customer churn prediction in order to detect the customers that are likely to churn i.e. switch from one service provider to another. Several data mining techniques are applied for classifying customers into the churn and non-churn category. But churn prediction applications comprise an imbalanced distribution of the dataset.

One of the commonly used techniques to handle imbalanced data is re-sampling of data as it is independent of the classifier being used. In this paper, we develop a hybrid re-sampling approach named SOS-BUS by combining well known oversampling technique SMOTE with our novel under-sampling technique. Our methodology aims to focus on the necessary data of majority class and avoid their removal in order to overcome the limitation of random under-sampling. Experimental results show that the proposed approach outperforms the other reference techniques in terms of Area under ROC Curve (AUC).

Index Terms**—**Imbalanced data, Re-sampling, Under-sampling, Classifier ensemble, Churn prediction.

I. INTRODUCTION

In recent years, increasingly competitive environment and saturation in the market has raised the need to focus on customer retention. Companies are more concerned with retaining existing customers than acquiring new customers because the cost of acquisition is much higher than the cost of retention [1]. Also, long-term customers are beneficial in many respects [2]:

1. They are less likely to be affected by the competitor's activities.
2. Relatively less cost to serve them.

In turn, their retention is more advantageous to the company. As a result, companies are targeting the existing customers' satisfaction and taking some action for those who are not satisfied and likely to change the service provider. To meet this target, correct identification of unsatisfied customers is necessary. That is, existing customers are categorized into two groups, namely churn wherein customers are likely to change the service provider and non-churners that are not likely to change the provider. Classifying the customers into one of these categories correctly help to focus on the right customers. Hence, researchers are paying attention on making use of data mining techniques for churn prediction.

Churn prediction applications have skewed distribution of dataset wherein one class has very high number of instances compared to the other class. The class with fewer samples is minority class and the one with relatively extra instances is majority class. the imbalance between two classes is represented by using 'Imbalance ratio' which is defined as the ratio of a number of samples in the majority class to that of a minority class. In customer churn prediction, number of non-churners is relatively high with respect to the number of churners. The classifier that is trained on such skewed dataset may bias towards majority class in a decision-making process, i.e. most of the times minority instance is classified as the majority. Consider churn prediction system where customers are categorized as either churn or non-churn customer based on various attributes. Assume that, dataset has 6 churn customers and 94 non-churn customers and the classifier achieves 94% accuracy. Thus classifier that is trained with imbalanced data has classified all majority instances correctly, but minority instances incorrectly. However, we are actually interested in finding out churn instances correctly and hence accuracy of 94% is not useful for us as churn customers are classified as non-churners. That is, accurate detection of those rare churners is critical for retaining them.

Various approaches have been proposed to resolve problems associated with imbalanced data set. These approaches can be categorized into four categories [3, 4].

- Data level approaches
- Algorithm level approaches
- Cost-sensitive learning approaches
- Classifier Ensemble techniques

Data Level Approach (DLA):

Data level approaches [4] apply re-sampling technique to the imbalanced data in order to alter the distribution of the majority, minority or both the classes. Depending on the class that is being re-sampled, these re-sampling approaches are known as the under-sampling, oversampling or hybrid approach. Use of re-sampling will cause a reduction in the imbalance ratio, which will lead to the performance improvement of the classifier.

Algorithm Level Approach (ALA):

In this category, the existing learning algorithm is modified so that it is able to cope up with an imbalanced distribution of data. However, the extent of effectiveness of the learning algorithm depends on the choice of the learning algorithm.

Cost-Sensitive Learning Approach (CSLA):

Cost-sensitive learning approaches [3] make the joint use of data level and algorithm level approach which not only applies data level pre-processing but also assigns different misclassification costs to the majority and minority class. These methods assign more weights to the samples of minority class than samples of majority class [4]. This helps in enhancing the performance of classifier provided that proper cost matrix is decided.

Classifier Ensemble Approach (CEA):

The ensemble of classifiers is constructed by combining different individual classifiers that can jointly perform the classification task. The approach makes use of a number of different classifiers to classify a given instance and then combines the decisions of multiple classifiers to give the final decision [5]. Recent experimental studies show that classifier ensemble technique helps to improve classification performance provided multiple diverse classifiers are combined intelligently.

Above four approaches can be summarized as shown in Table 1.

Table 1. Approaches to Handle Class Imbalance Issue

Method	Characteristics	Advantages	Limitations
DLA	Uses pre-processing technique	Independent of the classifier being used	Can remove necessary data or cause over-fitting
ALA	Modify existing algorithm	Improves performance	Effectiveness depends on the choice of learning algorithm & problem domain
CSLA	Combination of data & algorithmic level	Avoids costly errors	Deciding appropriate cost matrix
CEA	Combine diverse set of classifiers	Improves performance compared to an individual classifier.	Identification of diverse base classifiers.

Data level approaches require less computational time for data preparation. That is, we may preprocess the dataset only once and can use it to train different classifiers. In addition to that, their use is independent of the classifier being used. Nowadays, researchers are paying attention to the usage of classifier ensemble as combining the opinions of many experts are likely to improve the probability of success.

To exploit the advantages of both, we propose a hybrid approach that combines re-sampling with a classifier ensemble technique. The hybrid re-sampling approach is applied to the original imbalanced data to reduce the imbalance between the classes. Then the re-sampled data is provided as input to train the classifier ensemble model. The decisions of individual classification algorithms are combined to get the final decision of the class to which the given instance belongs.

The rest of this paper is organized as follows. Section 2 presents a brief overview of related work. In section 3, we discuss different re-sampling techniques for imbalanced datasets. Section 4 describes the proposed methodology. Experimental setup and evaluation parameters are discussed in section 5 followed by results and discussions in section 6. Finally, section 7 concludes the paper.

II. Related Work

Wouter Verbeke et al. [2] proposed two rule induction techniques AntMiner+ and Active Learning Based Approach for SVM rule extraction (ALBA) that can generate correct and intelligent rule sets for classification. AntMiner+ is based on the principles of Ant Colony and uses MAX-MIN Ant system. Though it generates less sensitive rule sets, their size is smaller and can incorporate domain knowledge. ALBA involves active learning wherein emphasis is given to the problem areas having higher noise. When it is combined with Ripper or C4.5, it gives the highest accuracy.

J. Burez et al. [6] investigated the impact of random under-sampling and advanced under-sampling technique CUBE on the class imbalance in six real-life customer churn prediction data sets. The results show that usage of advanced sampling technique CUBE has not proved much beneficial. The authors also implemented two modeling techniques, namely gradient boosting and weighted random forests and demonstrated performance improvement in terms of AUC and lift relative to standard techniques such as Logistic regression and Random Forest.

Wouter Verbeke et al. [7] developed a new evaluation measure that aims to maximize the profit. Author's findings suggest that data quality plays a crucial role in the classifier performance and six to eight attributes can find churn customers with high accuracy. Hence it is better to focus on collecting good quality data than collecting a huge amount of data.

Kyoungok Kim et al. [8] presented a churn prediction method that uses not only customer's personal information but also communication patterns among customers as it may affect churn. For this, a new variable known as network variable is derived from network analysis and incorporated in the training dataset. A measure called as Eigenvector centrality that indicates values of the first Eigenvector of the graph adjacency matrix are used to detect the customers who are likely to influence more on other customers. The authors suggest the use of another available centrality measure for this purpose.

T. Vafeiadis et al. [9] presented a comparative study of five widely used algorithms, namely Artificial Neural Network, Support Vector Machines, Decision Trees learning, Naïve Bayes and Regression analysis logistic regression analysis for churn prediction systems. Those classifiers were also applied with boosting and except Naïve Bayes and Logistic Regression, have shown accuracy improvement of 1 – 4 % and F-measure improvement of 4.5 – 15%.

M. Alper Tunga et al. [10] introduced a polynomial based method Euclidean Indexing High Dimensional Model Representation (HDMR) for churn prediction. Initially, attribute filter 'Information gain' is applied to select necessary attributes. Experimental evaluation in terms of accuracy, Sensitivity, specificity, precision, and RMSE prove significant performance improvement with the help of the proposed method. The authors conclude that proposed approach is more efficient and reliable.

Adnan Amin et al. [11] used rough set theory as a basis to extract the decision rules. Four rule generation algorithms namely, Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA) and the LEM2 were experimentally evaluated in terms of precision, recall, the rate of misclassification, lift, coverage, accuracy, and F-measure. The authors suggest RST approach using GA outperforms other rule generation techniques. Wenjie Bi et al. [12] proposed a new clustering algorithm called Semantic Driven Subtractive Clustering Method (SDSCM) that provides effective data analysis through its parallel implementation. It also increases the accuracy and decreases the risk of inaccurate operations.

Ning Lu et al. [13] presented boosting based technique that makes use of weights to segregate the customers into two clusters. One of those clusters will identify the high churn risk group. A separate classification model is built for each cluster and evaluated using AUC and lift curve as an evaluation measure. The method initially applies Chi-square Automatic Interaction Detection (CHAID) analysis in order to select the attributes which, when combined, will give the best predictive model and remove irrelevant attributes. Experimental results show better performance gains for churn prediction with the proposed approach.

Wai-Ho Au et al. [14] proposed a novel evolutionary algorithm data mining by evolutionary learning (DMEL) that not only predicts whether the customer will churn but also the likelihood of doing so. The scheme is able to handle the missing values in an efficient manner. Experimental results on the proposed approach prove its capacity to find hidden regularities in the database and outperform the techniques like neural networks and C4.5. It is robust in nature and can predict the churns under different churn rates.

Bing Zhu et al. [15] presented an experimental study of various techniques designed to handle class imbalance in churn prediction. Performance evaluation of 21 different techniques in terms of AUC, lift and profit based measure EMP shows the significant impact of evaluation measures on the performance. The authors conclude that classifier ensemble techniques are prominent in handling imbalanced data effectively. Also bagging and random forest algorithms generate best results for a profit-based measure.

Thomas Verbraken et al. [16] introduced a novel probabilistic metric EMPC (Expected Maximum Profit Measure for Customer Churn) in order to maximize the profit. Though AUC is a good measure for classification performance, it is based on some assumptions about the misclassification costs and may not prove beneficial in business environments. EMPC not only incorporates losses and gain but also decides the set of customers that should be focused in order to maximize the profit. Sensitivity analysis of a proposed measure shows its robustness towards changes in the parameters.

Ammar A. Q. Ahmed et al. [17] presented a meta-heuristic based churn prediction technique for the huge telecom dataset. Firefly algorithm can efficiently handle churn dataset but is computationally intensive. The proposed hybrid method overcomes this limitation with the help of simulated annealing and speeds up the classification process. The analysis in terms of ROC, PR, F-Measure, Accuracy and Time proves the performance of the proposed approach as good as the firefly algorithm with low time complexity.

Adnan Amin et al. [18] presented a study of six well-known sampling techniques , Mega-trend Diffusion Function (MTDF), Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling approach (ADASYN), Couples Top-N Reverse k-Nearest Neighbor (TRkNN), Majority Weighted Minority Oversampling Technique (MWMOTE) and Immune centroids oversampling technique (ICOTE) on imbalanced datasets. Experimental results show the best classification performance can be achieved by making combined use of MTDF for re-sampling and Genetic algorithm for rule generation.

III. Data Level Pre-processing

A. Oversampling

Oversampling technique reduces the imbalance ratio of skewed data set by duplicating minority instances [19]. This retains all the existing instances of the training data set but causes an increase in the size of the data set [20]. This balancing of data set helps to improve the performance of classification algorithm but can cause the over-fitting problem. That is, generated decision regions are very specific and can degrade the performance of a classifier. Hence, most commonly used an oversampling technique known as the Synthetic Minority Over-sampling Technique (SMOTE), introduces additional synthetic samples to the minority class rather than duplicating the instances directly.

Synthetic Minority Oversampling (SMOTE)

SMOTE approach [21] introduces synthetic examples in order to increase the instances of a minority class. The idea aims to overcome the limitation of over-sampling with replacement where specific decision regions are created. Use of synthetic samples helps to create larger and less specific decision regions. The algorithm first finds k nearest neighbors of each minority class sample by using Euclidean distance as a distance measure. Synthetic examples are generated along the line segments joining original minority class sample with its k nearest neighbors. The value of k depends on a number of artificial instances need to be added.

Steps for generation of synthetic samples:

1. Generate a random number between 0 and 1.
2. Calculate the difference between the feature vector of minority class sample and its nearest neighbor.
3. Multiply this difference by a random number generated in step 1.
4. Add the result of multiplication to the feature vector of minority class sample.
5. The resulting feature vector identifies the newly generated sample.

B. Under-sampling

The under-sampling technique reduces the imbalance ratio of the skewed dataset by removing some samples from the majority class [19]. It decreases the number of samples of majority class in order to make the number of instances of two classes approximately equal. The produced balanced dataset will help to improve the performance of learning algorithm. But as the instances to be deleted are selected randomly from the majority class, the approach can sometimes remove necessary and important instances from the training dataset. This attempt to achieve generalization may result in a less representative distribution of the training data [20] which in turn may degrade the classifier performance.

C. Hybrid Approach

The hybrid approach uses the combination of over-sampling and under-sampling in order to balance the dataset, i.e. some of the majority class instances will be deleted and some instances will be added to the minority class. The previous work shows that combined use of oversampling and under-sampling results in the performance improvement compared to the performance of the individual technique. The representatives of this category are BorderLine-SMOTE1, BorderLine-SMOTE2, SMOTE-Tomek Links, SMOTE-ENN, and Safe-Level-SMOTE.

a. BorderLine-SMOTE1 and BorderLine-SMOTE2

Borderline-SMOTE1 and Borderline-SMOTE2 techniques [22] enhance SMOTE in order to focus on only borderline instances of the class. These approaches apply oversampling to only minority class examples that are near the borderline. Initially, all the instances of minority class are categorized into one of the three categories, namely noise, safe or dangerous. Initially, it identifies the Borderline instances of a minority class. Then synthetic instances are generated along the line between borderline instances and their selected nearest neighbors. In the case of Borderline-SMOTE1, the selected nearest neighbors are only from the minority class while in the case of Borderline-SMOTE2 they can be of the majority class also.

b. SMOTE-Tomek Links

This hybrid approach [23] is introduced in order to handle the situation where class clusters are not well defined due to the existence of some majority class examples in the minority class space. To resolve this issue, Tomek Links are used as a data cleaning method. Tomek link can be defined as [24]:

If two instances S_1 and S_2 belong to different classes and d (S_1, S_2) is the distance between S_1 and S_2

Then (S_1, S_2) pair is called as Tomek link if there does not exist the example S_3 that satisfies the condition

$$d(S_1, S_3) < d(S_1, S_2) \tag{1}$$

or

$$d(S_2, S_3) < d(S_1, S_2) \tag{2}$$

If two examples form a Tomek link, then either one of these examples is noise or both examples are borderline examples. In this approach, Tomek links are applied to the over-sampled training set for the purpose of cleaning the data.

c. SMOTE-ENN

Edited Nearest Neighbor (ENN) Rule [23] applies data cleaning by removing examples from both the classes. ENN removes more instances than those removed by

Tomek Link. The examples to be removed are selected based on the number of nearest neighbors misclassifying it. That is, if three nearest neighbors of any instance misclassify the instance, then it is removed from the training set.

IV. Proposed Methodology

The proposed work is carried out in two phases:

1. A Novel Hybrid Approach: SMOTE Oversampling and Borderline Under-sampling (SOS - BUS)
2. Classifier Ensemble Formation

A. A Novel Hybrid Approach: SMOTE Oversampling and Borderline Under-sampling (SOS - BUS)

SMOTE [3], a well-known over-sampling technique is applied to the original skewed data set in order to introduce the synthetic samples to the minority class. This will increase the number of samples of the minority class which in turn reduces the imbalance ratio of the dataset.

Random under-sampling is a data level pre-processing approach that tries to balance imbalanced dataset by deleting some of the majority class samples. The instances to be deleted are selected randomly from the original dataset which raises one major issue that needs attention. Random selection of instances can sometimes remove necessary instances of majority class [25]. Our proposed work emphasizes on this issue and tries to maintain necessary instances of majority class.

Instances which are lying on the borderline of two classes are necessary instances of the class as they are critical in deciding the decision boundary and their removal can degrade the performance of the classifier. Hence proposed methodology initially identifies the borderline instances of majority class and adds them to the output dataset. Then random under-sampling is applied to the left over data which will randomly remove some of them in order to balance the dataset. This ensures that necessary data is not removed and performance is not degraded. Fig. 1 represents the methodology as a whole.

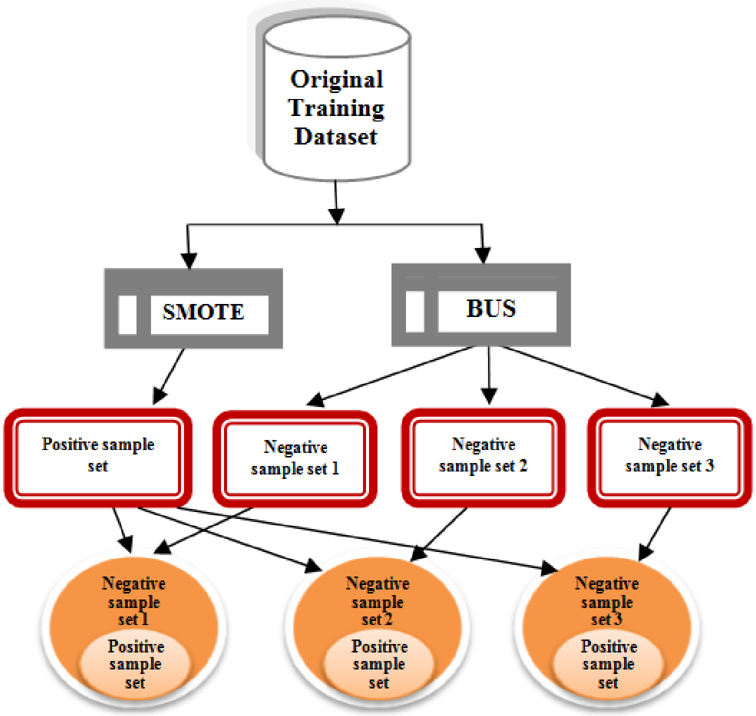


Fig.1. SOS-BUS Re-sampling Technique

Steps of the SOS - BUS Methodology

Input:

T_{imbal} : Input Training data set
B : Set of majority class instances
S : Set of minority class instances
B_N : Set of k nearest neighbors of instance of majority class
B_{MN} : Set of nearest neighbors that belongs to minority class
B_{border} : Set of borderline neighbors of majority class
T_{bal} : Output Training data set after pre-processing
bnum :Number of majority class elements
snum :Number of minority class elements

1. *Oversampling using SMOTE*

Apply Synthetic Minority Oversampling pre-processing technique to dataset T_{imbal} in order to introduce artificial instances in the minority class and generate data set T'_{imbal}.

2. *Find k nearest neighbors of each sample S*

Calculate Euclidean distance d_1 between sample q and other instance T_1 of the training dataset. Store the distance d1 in array Dist. Repeat this process for all other instances T_i so that the array will contain distance d_1,d_2…..dn between sample q and all other instances $T_1,T_2,\ldots,T_n$. Sort this array in the increasing order of distances and select first k instances as nearest neighbors of sample q.

For every point q in majority class B, find the set of k points B_N such that,
For all B_i Є B_N and for all B_j ₵ B_N

$$d(B_i,q) \le d(B_j,q) \qquad (3)$$

where
Bi: ith instance of set B
q: an instance of set B.

3. *Identification & Removal of noisy instances*

Identify noisy instances and delete them from the dataset T'_{imbal}.
Noisy instance: Instance q whose class differs from the class of all k nearest neighbors of the sample q.

4. *Identification of necessary data of majority class*

Identify borderline instances B_{border} of majority class and push them to the output data set T_{bal} because these are the necessary instances of majority class.

For every point q in majority class B, find the set of k' points B_{MN} such that,

$$B_{MN} = \{x \mid x \in B_N \wedge class(x) \neq majority\} \qquad (4)$$

Find the set of points B_{border} such that, for every B_i Є B_{border} there exists set B_{MN} that satisfies the following condition

$$k/2 < |B_{MN}| < k \qquad (5)$$

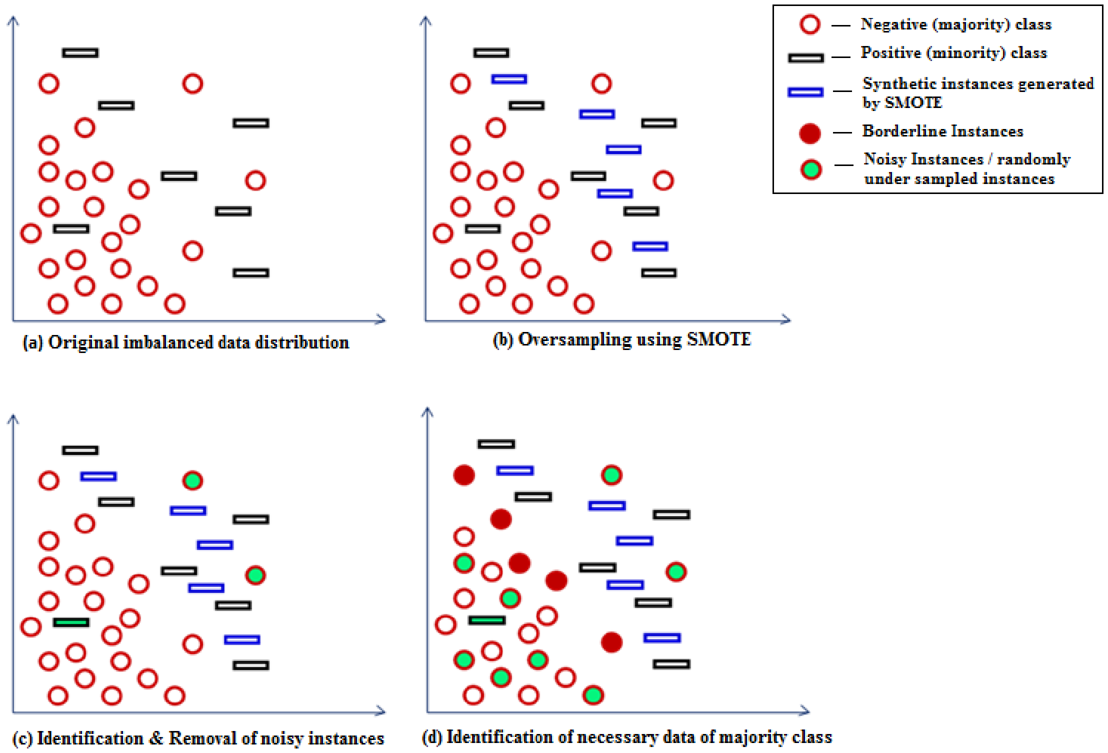
where k indicates a number of nearest neighbors selected.

5. *Under-sampling of majority data*

Consider data set

$$E = T'_{imbal} - B_{border} \qquad (6)$$

Apply random under-sampling to data set E and push randomly selected samples to the output data set T_{bal}. This output data set T_{bal} is in balanced form and is used for classification.
Fig. 2 represents the stepwise procedure of the proposed approach


Negative (majority) class
Positive (minority) class
Synthetic instances generated by SMOTE
Borderline Instances
Noisy Instances / randomly under sampled instances

(a) Original imbalanced data distribution

(b) Oversampling using SMOTE

(c) Identification & Removal of noisy instances

(d) Identification of necessary data of majority class

Fig.2. Stepwise SOS – BUS Approach

Fig. 2a shows the original imbalanced distribution of instances of two classes. Initially, popular oversampling technique SMOTE is applied to skewed data so that representative instances of minority class are increased by adding synthetic instances to it. Blue rectangles in Fig. 2b represent the synthetic instances introduced with the help of SMOTE. Fig. 2c highlights the shapes filled with green color to point out noisy instances of the dataset. The instances that are surrounded by all the instances of another class are considered as noise. Finally, necessary data of majority class, i.e. borderline instances of majority class are identified and added to the output set as they must be retained in the balanced dataset. Circles filled with Red color in Fig. 2d are the borderline instances of majority class and will surely be part of the output data set after pre-processing. Now Random Under-sampling is applied to the remaining instances of majority class. Though random under-sampling is applied, it may not remove necessary data of the majority class because we have already identified that data and such necessary data won't be applied to random under-sampling.

B. *Classifier Ensemble Formation*

To construct the classifier ensemble, diversity is introduced with the help of different training subsets. The subset generation process starts by initially applying SMOTE to the minority class. Then our proposed under-sampling algorithm (BUS) is applied to the majority class. Its output is combined with the oversampled minority class to get one training subset. The process is repeated thrice to get three different subsets of training data. Every time our proposed BUS method will select the different subset of samples, hence those subsets will be different and will introduce diversity. Two base classifiers namely J48 and SVM are trained on these three training subsets. Predictions of those classifier models are combined using the voting method. Fig. 3 represents the classifier ensemble construction process.

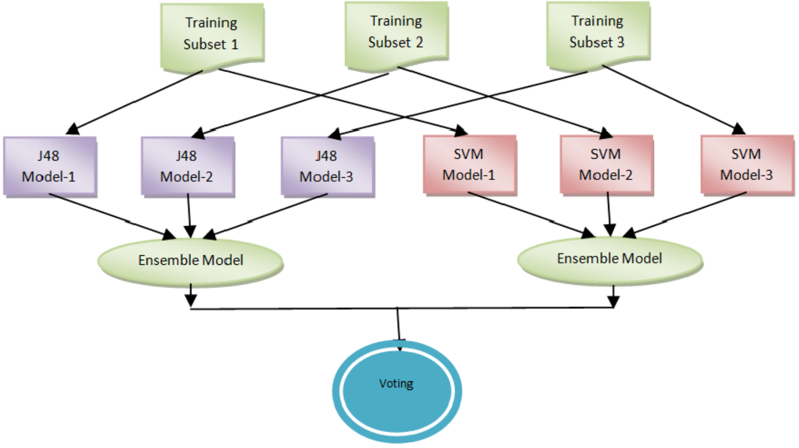


Fig.3. Classifier Ensemble Formation

V. Experimental Design

A. *Experimental Setup*

The experiments were carried out using Weka environment with its default parameters. Weka is an open source [26, 27] toolkit that provides a set of machine learning & pre-processing algorithms.

In this work, we implement a hybrid re-sampling approach to imbalanced data. Oversampling is done with the commonly used SMOTE while our novel BUS under-sampling technique reduces the samples of majority class. Then preprocessed data is provided as input to an ensemble of classifiers that is constructed by using different training datasets as well as different classifier models. Different classifier models, namely J48 and SVM are used and their results are combined with a voting method. All experiments were carried out 5 times using 10 fold cross validation.

B. *Experimental Data sets*

For experimental evaluation, we have used churn dataset from the UCI Machine Learning repository that is publicly available. It comprises data of 3333 telecom customer records with 21 features for each record. Out of those records, 483 customers are of churn class while 2850 belong to the non-churn class. For experimental evaluation, subsets of this dataset with varying imbalance ratios are selected randomly.

C. *Evaluation Parameters*

Existing classification systems have been evaluated by using various parameters such as accuracy, F-measure, type-I error, type-II error, G-mean, AUC (Area under ROC curve) etc. These parameters can be derived with the help of a confusion matrix. Table 2 presents a confusion matrix [28] for a binary class that indicates correct and incorrect classifications.

Table 2. Confusion Matrix for a Two Class Problem

	Predicted as Churn	Predicted as Non-churn
Churn Class	True Positive	False Negative
Non-churn Class	False Positive	True Negative

Previous studies [29, 30] show that accuracy is not an appropriate measure to evaluate the performance of classifying imbalanced dataset. This is because it only considers a number of correctly classified instances but does not pay attention to how many of them are of a positive class which is important for us. Hence Area under the ROC curve (AUC) has been suggested as an appropriate measure. In this paper, we have used AUC as evaluation parameter which is defined as an arithmetic average of the mean predictions for each class [31]. AUC can be represented as

$$AUC = \frac{sensitivity + specificity}{2} \tag{7}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$specificity = \frac{TN}{FP + TN} \tag{9}$$

VI. Results and Discussion

Table 3 shows the experimental results in terms of AUC when different types of re-sampling techniques are applied to the imbalanced dataset which is to be classified. Initially, classification is done without applying any re-sampling techniques, i.e. data in the imbalanced form are provided as input and the results of classification performance in terms of AUC are recorded. Then well known oversampling technique SMOTE is used for pre-processing and results show that there is a significant improvement in the AUC value than the previous one. We implemented hybrid approach SOS – RUS that combines SMOTE with Random Under-sampling and observed further improvements in AUC compared to previous two approaches. Finally, we combined SMOTE with our proposed novel under-sampling approach (SOS – BUS) and got the higher AUC values than all the above approaches.

The experiments are carried out for varying imbalance ratios (IR) that indicate the ratio of a number of majority class instances to that of a minority class.

Fig. 4 plots performance comparison in terms of AUC when different re-sampling methods are applied to datasets whose imbalance ratios vary in the range of 1 to 120.

Fig. 5 plots performance comparison in terms of F-measure when different re-sampling methods are applied to datasets whose imbalance ratios vary in the range of 1 to 120.

Table 3. Values of AUC when different re-sampling methods are applied to churn data set

Dataset	IR	Re-sampling Method			
		None	SMOTE	SOS-RUS	SOS –BUS
Test1	1	0.828	0.865	0.854	0.858
Test2	5	0.833	0.878	0.879	0.885
Test3	10	0.824	0.885	0.844	0.863
Test4	15	0.827	0.868	0.855	0.875
Test5	20	0.842	0.86	0.86	0.881
Test6	25	0.846	0.863	0.862	0.88
Test7	30	0.851	0.859	0.812	0.878
Test8	35	0.802	0.852	0.881	0.89
Test9	40	0.787	0.846	0.836	0.861
Test10	45	0.786	0.801	0.798	0.821
Test11	50	0.82	0.844	0.818	0.857
Test12	55	0.798	0.808	0.803	0.822
Test13	60	0.769	0.788	0.743	0.802
Test14	80	0.756	0.806	0.777	0.825
Test15	100	0.72	0.741	0.742	0.799
Test16	120	0.628	0.72	0.654	0.787

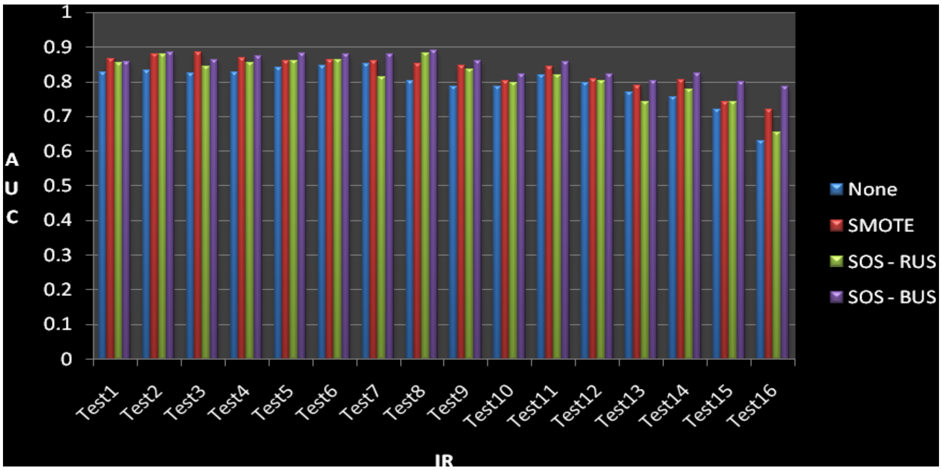


Fig.4. AUC for datasets of different IR

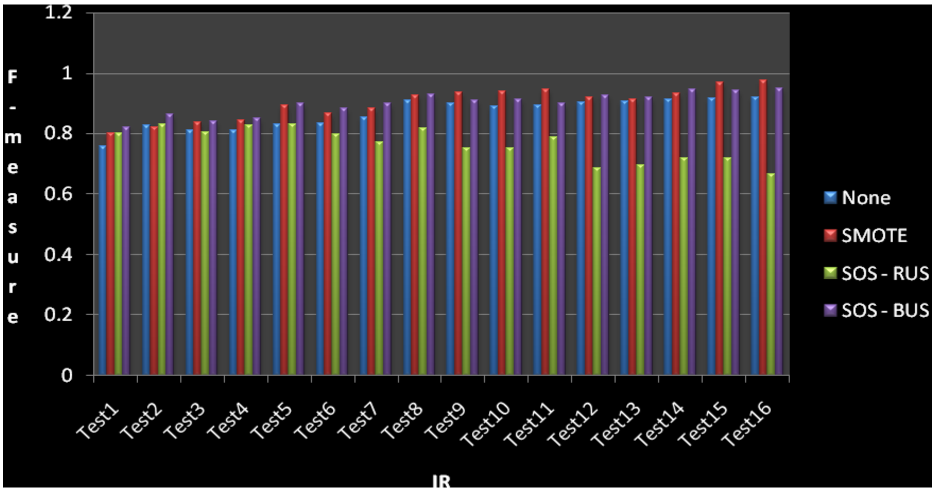


Fig.5. F-measure for datasets of different IR

Experimental work and analysis of above graphs lead to some interesting conclusions that are listed below:

- Above graphs clearly, show that in almost all cases the proposed pre-processing method SOS – BUS that combines SMOTE with proposed novel under-sampling technique results in the improvement of AUC. Although the improvement in AUC value seems relatively small, it is beneficial for skewed churn data set where correct identification of the churn class instance is extremely important.
- Initially, for lower imbalance ratio performance of SOS – RUS and that of SOS – BUS is nearly same and performance gains of the proposed methodology are not significant. But as IR increases, the differences start highlighting and benefits of the proposed method are prominent.
- In some of the observations, performance after applying SMOTE is better than that of SOS – RUS. This shows the chances that some important data might have been removed due to random under-sampling and this might have caused performance degradation. This issue has been resolved by the proposed approach.

VII. Conclusion

Churn prediction is a critical issue for many companies to be in the market competition. But churn datasets suffer from an imbalanced distribution of data and classifiers that are trained with such datasets may bias towards the majority class and degrade the classification performance. Hence dealing with imbalanced datasets is gaining the attention of many researchers in the area of pattern recognition. In this work, we have proposed a hybrid approach SOS – BUS that makes combined use of over-sampling and under-sampling techniques. Over-sampling is done by using a popular approach known as SMOTE while a novel approach is proposed for under-sampling in order to overcome the limitation of random under-sampling. The proposed approach focuses on borderline instances of majority class as these are considered as important data of the class. This is because those elements are critical in finalizing the decision boundary between the two classes. Experimental results of the proposed approach are compared with results of classifier without any pre-processing, classifier with commonly used over-sampling technique SMOTE and classifier with another hybrid approach SOS – RUS. Experimental results prove that the proposed approach outperforms other reference techniques.

References

[1] S. Y. Hung, D. C. Yen, and H. Y. Wang, “Applying data mining to telecom churn management,” *Expert Systems with Applications*, vol. 31, no. 3, pp. 515-524, 2006.

[2] W. Verbeke, D. Martens, C. Mues, and B. Baesens, “Building comprehensible customer churn prediction models with advanced rule induction techniques,” *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354-2364, 2011.

[3] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463-484, 2012.

[4] Y. Park, and J. Ghosh, “Ensembles of $({\ alpha}) $-Trees for Imbalanced Classification Problems,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 131-143, 2014.

[5] P. Cao, J. Yang, W. Li, D. Zhao, and O. Zaiane, “Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD,” *Computerized Medical Imaging and Graphics*, vol. 38, no. 3, pp. 137-150, 2014.

[6] J. Burez, and D. Van den Poel, “Handling class imbalance in customer churn prediction,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626-4636, 2009.

[7] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, “New insights into churn prediction in the telecommunication sector: A profit driven data mining approach,”. *European Journal of Operational Research*, vol. 218, no. 1, pp. 211-229, 2012.

[8] K. Kim, C. H. Jun, and J. Lee, “Improved churn prediction in telecommunication industry by analyzing a

large network," *Expert Systems with Applications*, vol. 41, no. 15, pp. 6575-6584, 2014.

[9] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1-9, 2015.

[10] M. A. Tunga, and A. Karahoca, "Detecting GSM churners by using Euclidean Indexing HDMR," *Applied Soft Computing*, vol. 27, pp. 38-46, 2015.

[11] A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain, and K. Huang, "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, pp. 242-254, 2017.

[12] W. Bi, M. Cai, M. Liu, and G. Li, "A big data clustering algorithm for mitigating the risk of customer churn," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1270-1281, 2016.

[13] N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1659-1665, 2014.

[14] W. H. Au, K. C. Chan, and X. Yao, "A novel evolutionary data mining algorithm with applications to churn prediction," *IEEE transactions on evolutionary computation*, vol. 7, no. 6, pp. 532-545, 2003.

[15] B. Zhu, B. Baesens, and S. K. vanden Broucke, "An empirical comparison of techniques for the class imbalance problem in churn prediction," *Information Sciences*, vol. 408, pp. 84-99, 2017.

[16] T. Verbraken, W. Verbeke, and B. Baesens, "A novel profit maximizing metric for measuring classification performance of customer churn prediction models," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 5, pp. 961-973, 2013.

[17] A. A. Ahmed, and D. Maheswari, "Churn prediction on huge telecom data using hybrid firefly based classification," *Egyptian Informatics Journal*, 2017.

[18] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access*, vol. 4, pp. 7940-7957, 2016.

[19] S. Hu, Y. Liang, L. Ma, and Y. He, "October. MSMOTE: improving classification performance when training data is imbalanced," In *Computer Science and Engineering, 2009. WCSE'09. Second International Workshop on* , vol. 2, pp. 13-17, IEEE, 2009.

[20] H. Cao, V. Y. Tan, and J. Z. Pang, "A parsimonious mixture of Gaussian trees model for oversampling in imbalanced and multimodal time-series classification," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 12, pp. 2226-2239, 2014.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.

[22] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," *Advances in intelligent computing*, pp. 878-887, 2005.

[23] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 20-29, 2004.

[24] I. Tomek, "Two modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, vol. 6, pp. 769-772, 1976.

[25] U. R. Salunkhe, and S. N. Mali, "Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach," *Procedia Computer Science*, vol. 85, pp. 725-732, 2016.

[26] G. Wang, J. Ma, and S. Yang, "An improved boosting based on feature selection for corporate bankruptcy prediction," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2353-2361, 2014.

[27] Doaa Hassan,"The Impact of False Negative Cost on the Performance of Cost Sensitive Learning Based on Bayes Minimum Risk: A Case Study in Detecting Fraudulent Transactions", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.2, pp.18-24, 2017. DOI: 10.5815/ijisa.2017.02.03

[28] C.Bhanuprakash, Y.S. Nijagunarya, M.A. Jayaram,"Clustering of Faculty by Evaluating their Appraisal Performance by using Feed Forward Neural Network Approach", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.3, pp.34-40, 2017. DOI: 10.5815/ijisa.2017.03.05

[29] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539-550, 2009.

[30] L. Abdi, and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 238-251, 2016.

[31] A. I. Marqu és, V. García, and J. S. Sánchez, "Two-level classifier ensembles for credit risk assessment," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10916-10922, 2012.

Authors' Profiles

Uma R. Salunkhe has completed her Master of Engineering in CSE-IT from the University of Pune. She is research scholar at Smt. Kashibai Navale College of Engineering, Pune and working as an Assistant Professor at Sinhgad College of Engineering, Pune, Maharashtra, India. She has 15 years of experience in teaching field. She has published 10 articles in various international journals and conferences. Her area of interest includes Security in Networks and Machine Learning.

Suresh N. Mali has completed his Ph.D. in Computer Science from Bharati Vidyapeeth, Pune and presently he is working as Principal in Sinhgad Institute of Technology and Science, Narhe, Pune, Maharashtra, India. He is the author of 3 books and has more than 40 research papers in referred international and national journals and conferences. His research interests mainly include Image Processing, Security, and Machine Learning.

How to cite this paper: Uma R. Salunkhe, Suresh N. Mali, "A Hybrid Approach for Class Imbalance Problem in Customer Churn Prediction: A Novel Extension to Under-sampling", International Journal of Intelligent Systems and Applications(IJISA), Vol.10, No.5, pp.71-81, 2018. DOI: 10.5815/ijisa.2018.05.08