Hybrid Approach Redefinition (HAR) model for optimizing hybrid ensembles in handling class imbalance: a review and research framework

Hartono Hartono[1,2,*], *Opim Salim* Sitompul[2], *Tulus* Tulus[2], *Erna Budhiarti* Nababan[2], and *Darmawan* Napitupulu[3]

[1]STMIK IBBI, Department of Computer Science, Medan, Indonesia
[2]Universitas Sumatera Utara, Department of Computer Science, Medan, Indonesia
[2]Lembaga Ilmu Pengetahuan Indonesia, Jakarta, Indonesia

Abstract. The purpose of this research is to develop a research framework to optimize the results of hybrid ensembles in handling class imbalance issues. The imbalance class is a state in which the classification results give the number of instances in a class much larger than the number of instances in the other class. In machine learning, this problem can reduce the prediction accuracy and also reduce the quality of the resulting decisions. One of the most popular methods of dealing with class imbalance is the method of ensemble learning. Hybrid Ensembles is an ensemble learning method approach that combines the use of bagging and boosting. Optimization of Hybrid Ensembles is done with the intent to reduce the number of classifier and also obtain better data diversity. Based on an iterative methodology, we review, analyze, and synthesize the current state of the literature and propose a completely new research framework for optimizing Hybrid Ensembles. In doing so, we propose a new taxonomy in ensemble learning that yields a new approach of sampling-based Ensembles and will propose an optimization Hybrid Ensembles using Hybrid Approach Redefinition (HAR) Model that combines the use of Hybrid Ensembles and Sampling Based Ensembles methods. We further provide an empirical analysis of the reviewed literature and emphasize the benefits that can be achieved by optimizing Hybrid Ensembles.

1 Introduction

Class imbalance occurs when the classification process gives results where there is a class with a number of instances much higher than the other classes. This problem generates majority and minority class issues. This problem causes the pattern contained in the minority class to be neglected. In machine learning and pattern recognition, the patterns contained in the minority class are important enough to be noticed because they contain unusual behaviors that distinguish them from general access patterns [1]. Hartono *et al.* has proposed a method for determining the centroid of K-Means that can minimize the class imbalance problem [2]. In other cases, such as clustering, the focus given is on the majority class because it tends to affect the accuracy of the clustering results. In this case better predictive results will be given if a class has a large number of instances, while classes with few instances will have poor predictive results [3]. The research of Hartono *et al.* shows that the methods in overcoming class imbalance problem need to pay attention to diversity data [4]. In addition, class imbalance issues can also generate bias in the decision-making process when decisions are tended to focus on majority class [5].

Alibeigi *et al.* groups approaches for handling class imbalance problems into four approaches: Feature Selection Approach, Level Approach Algorithm, Cost Sensitive Approach, and Ensemble Learning method [6]. Galar *et al.* Develop an ensemble learning based approach that combines a number of classifier to obtain single classifier. Ensemble learning methods are divided into 4 approaches arranged in the form of taxonomy, which consists of: cost-sensitive boosting, boosting-based ensembles, bagging-based ensembles, and hybrid ensembles [3]. Hybrid ensembles are an approach that combines the use of preprocessing stages with processing stages which are ensemble learning methods by combining the use of bagging and boosting, where in a number of stages of bagging a number of boosting processes are performed.

Galar *et al.* using UnderBagging and AdaBoost methods as processing steps in the proposed Hybrid Ensembles method and the research shows that the use of UnderBagging and AdaBoost methods requires a large number of classifier and also poorly generated data diversity. The preprocessing method used in this research is the SMOTEBoost method [3]. The taxonomy put forward by Galar *et al.* has been used extensively by a number of researchers. Galar *et al.* tried to throw taxonomy using the new Boosting approach using EUSBoost but the data diversity obtained was not good enough [7].

* Corresponding author: hartonoibbi@gmail.com

2 Related Works

Fernandez *et al.* uses hybrid ensembles in dealing with multi-class problems that combine the use of SVM Methods in making multi-class problems into two classes, but the resulting accuracy is low [8]. Zaki & Meira in a study conducted using the Sample Subset Optimization method and gave the result that the Hybrid Ensembles method should pay attention to the Diversity Data problem [9].

The preprocessing method includes a number of stages such as: data extraction, data cleaning, data fusion, data reduction, and feature construction [9]. The preprocessing method can be done through the process of discretization of numeric attributes, subset attribute selection, and handling missing value [10]. Studies conducted by Krawczyk modified hybrid ensemble methods and incorporated feature selection processes at the bagging stage by using a genetic algorithm, but the method they proposed failed to address the problem of data accuracy and their results showed that diversity measurements need attention [11].

The sampling-based Ensembles method began to develop since 2015, where Jose *et al.* proposed the Random Balance Ensemble Method that integrates random undersampling with SMOTEBoost [12]. According to Jian *et al.*, the Random Balance Ensemble Method method can also be used as a preprocessing stage, where the stage itself is intended to reduce the size of training data and the diversity of data and then this research presents a new method of ensemble learning called Different Contribution Sampling (DCS), which can be said to be a sampling-based and Boosting method [13]. The DCS method, the Biased Support Vector Machine (B-SVM) used to generate Non-Support Vector (NSV) and Support Vector (SV) combined with SMOTE (SV-SMOTE) is used to increase the number of members of the minority class and NSV combined RUSBoost (NSV-RUS) is used to decrease the number of members of the majority class. The use of DCS methods based on their research results can improve the diversity of data.

Tang & He proposed the GIR-Based Ensemble Sampling Approach method to overcome the problem of class imbalance [14]. GIR-Based Ensemble Sampling Approach combines the use of Undersampling and Oversampling used to improve the quality of single classifier in overcoming class imbalance problems and this research focuses on improving classifier quality and reducing classifier size. Ren *et al.* proposes a sampling-based ensemble learning method with an Ensemble Based Adaptive Over-Sampling method that modifies the Over-Sampling method using Adaptive SMOTEBoost in overcoming class imbalance problem and this Research also focuses on training classifier issues [15]. Research conducted by Gong & Kim proposes RHBoost (Random Hybrid Sampling Boosting) method which combines RUSBoost method with ROSE Sampling in overcoming class imbalance problems and gives better results compared to RUSBoost with a much smaller number of classifier and performance, but performance Which is obtained poorly for large datasets [16].

Further research on sampling-based ensemble learning was conducted by Lu *et al.* which proposed the Adaptive Ensemble Undersampling-Boost that incorporated the Ensemble of Undersampling (EUS), AdaBoost, and Cost-Sensitive Weight Modification methods for the class imbalance problem and provides good results for the number of classifiers but provides poor results for diversity data [17].

With regard to the taxonomic model proposed by Galar *et al.* [3], we propose a new taxonomic model that modifies the taxonomic model. We propose a taxonomy that yields a new approach of sampling-based Ensembles. We also put forward and apply the Hybrid Approach Redefinition (HAR) model that combines the use of Hybrid Ensembles and Sampling Based Ensembles methods. Where Random Balance Ensemble Method will be used for preprocessing stage while processing will be done by using DCS and Bagging method. Inside the DCS method itself is actually an ensemble of the sampling method with the Boosting method. There are 2 (two) major issues that will be answered that is about the problem of the number of a classifier and divesity data. The preprocessing method to be used is the Random Balance Ensemble Method [11]. The use of DCS methods coupled with UnderBagging is expected to reduce the number of classifiers and also increase the diversity of data [3].

3 Methodology

The Hybrid Ensembles method proposed by Galar *et al.* is a method that combines the bagging process with boosting [3]. In every stage of bagging will be done a number of boosting process. The Hybrid Ensembles method has been pretty good at handling the imbalance class but has a disadvantage in terms of data diversity.

3.1. Datasets

More than 40 dataset from KEEL data-set repository has been used for proposed method [19]. Multi-class data-sets are converted into two classes by grouping the class into positive classes and negative classes [3]. The selected data-sets reflect the varying imbalance ratio.

3.2. Research Process

The phase of the research process can be seen in Figure 1.

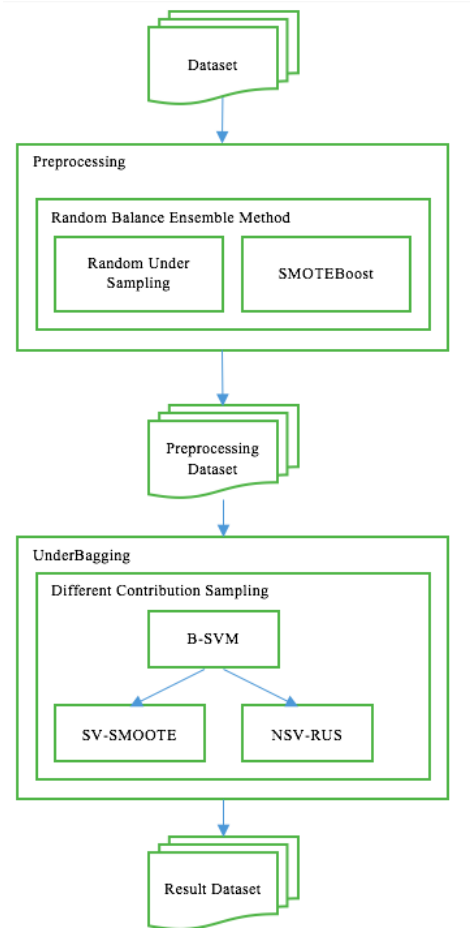

Fig. 1. Phases of the Research Process.

In Fig. 1, it can be seen that the dataset will be processed at the preprocessing stage. This preprocessing stage aims to reduce the number of a classifiers. This preprocessing stage will be done by using Random Balance method which is a combination of Random Undersampling and SMOTEBoost method. The result of this preprocessing step will result in a preprocessing dataset where the preprocessing dataset is smaller than the original dataset. The dataset generated from this preprocessing process subsequently undergoes processing using the HAR model. This HAR model consists of an Underbagging method in which each stage of bagging will be performed in a number of different stages of Different Contribution Sampling (DCS).

3.3. Random Balance Ensemble Method

Random Balance Ensemble Method stages in detail can be seen in Block Diagram as in Figure 2 [12].

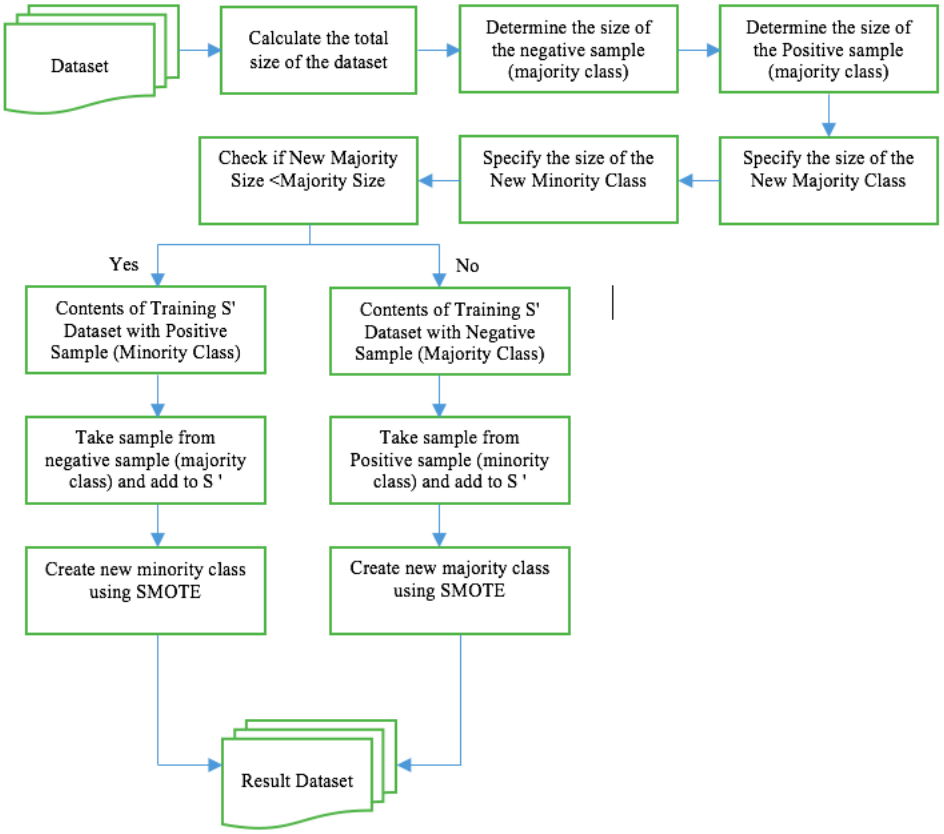

Fig. 2. Stages in the Random Balance Ensemble Method.

In Figure 2. it can be seen that the Random Balance Ensemble Method has a number of stages as follows.

1. the number of instances of the dataset and enter into the totalSize variable.
2. Take instance of majority class and insert into SN array variable, Majority Class is Negative Samples which will each instance in Negative Samples will be assigned value -1.
3. Take instance of minority class and insert into SP array variable. Minority Class is Positive Samples which each instance of Positive Samples will be +1.
4. Fill the value of the majoritySize variable with the number of instances in SN.
5. Fill the value of the minoritySize variable with the number of instances in SP.
6. Fill in the value of the newMajoritySize variable by generating a random number between 2 to the totalSize - 2 value.
7. Fill the value of the newMinoritySize variable with the totalSize value reduced by the newMajoritySize value.
8. If the value of newMajoritySize <MajoritySize then
   a. Contents of the Training S 'dataset variable with existing data on the SP array
   b. Perform SMOTEBoost process for data on Majority Class
   c. Eject the data on the Majority Class with the lowest proximity level and enter it into the Minority Class
9. If value newMajoritySize> MajoritySize then
   a. The contents of the Training S 'dataset variable with the data contained in the SN array
   b. Perform SMOTEBoost process for data on Minority Class
   c. Eject the data on the Minority Class with the lowest proximity level and enter it into the Majority Class

The data used are 44 data-sets sourced from KEEL data-set repository. Multi-class data-sets are converted into two classes by grouping the class into positive classes and negative classes [3]. The selected data-sets reflect the varying imbalance ratio.

3.4. Different Contribution Sampling

The working process of DCS method can be seen in Figure 3 [13].

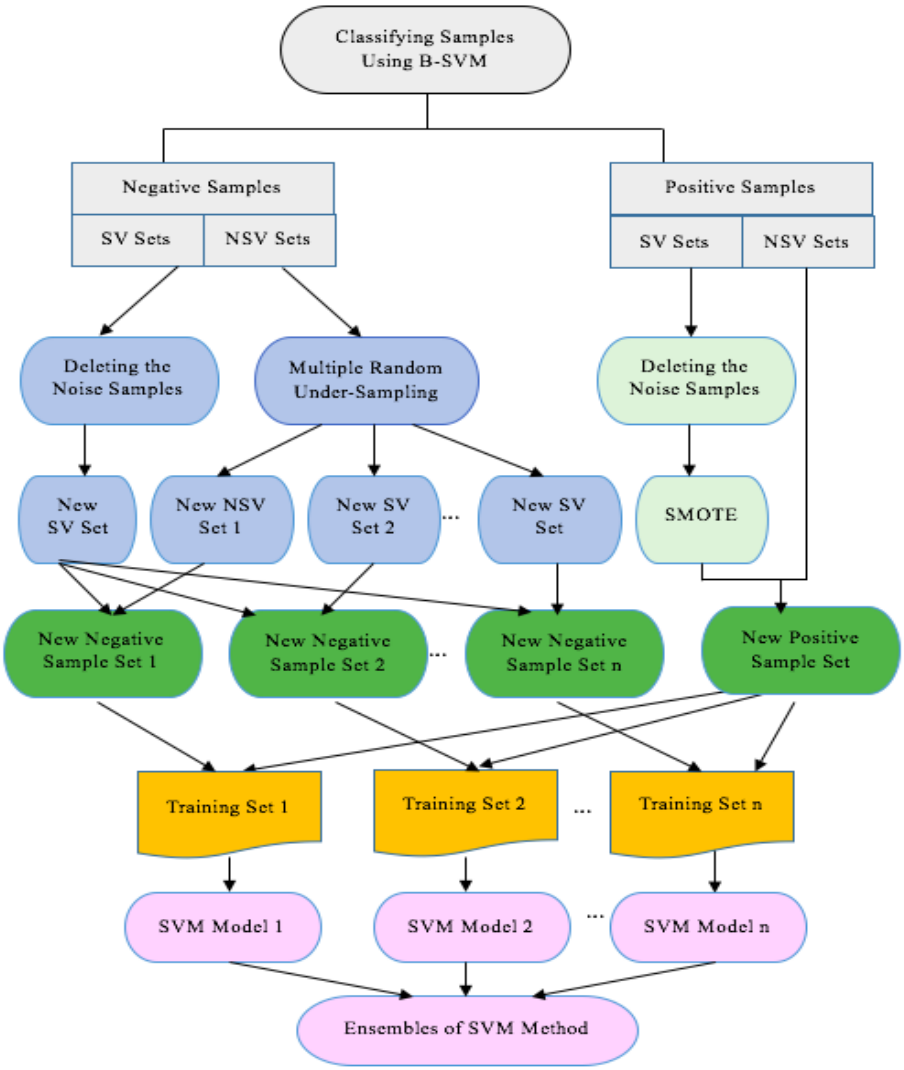

Fig. 3. Stages in the Different Contribution Sampling

In Figure 3. can be seen how the work method of Different Contribution Sampling (DCS) is covering a number of stages as follows.

1. Retrieve some data from minority class and majority class. Some data from the minority class are expressed as positive samples and some data from the majority class is expressed as negative samples.
2. Determine the hyperplane combination of positive samples and negative samples. Where the combined value of positive samples is 1 and the combined value of negative samples is -1.
3. Hyperplane is obtained by using Biased Support Vector Machine which is done by minimizing the margin value from positive samples and negative samples using (1) and (2) [13].

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}({w_1}^2 + {w_2}^2 + {w_3}^2 + {w_4}^2) \quad (1)$$

In (1). it can be seen that half of the hyperplane equation is obtained from positive samples and half of the hyperplane equations are obtained from negative samples.

As for (2). is an equation for the determination of Positive Samples and Negative Samples [12]

$$f(x) = sign\left(\sum_{i=1}^{l} y_i (w.x_i) + b\right) \quad (2)$$

If the sign function gives a result greater than 0 then it is included in the minority class and if the sign function gives a result smaller than 0 then it is included in the majority class.

4. Perform training for minority class and majority class based on hyperplane equation obtained. Training process can be done using (2).
5. The result for minority class must be positive and the result for majority class must be negative. If the training result for an instance in the minority class gives a negative result then move the instance into the majority class and if the training result for an instance in the majority class gives a positive result then move that instance into the minority class.
6. Retrieve some data back from minority class and make it as positive samples and retrieve some data back from majority class and make as negative samples.
7. Determine the hyperplane of each positive samples and negative samples using equations (1) and (2)
8. Perform training process on minority class. for the process of determining SV Sets and NSV Sets by using hyperplane from positive samples. If the calculation result gives then it is categorized into SV Sets and otherwise categorized as NSV Sets.
9. Do the training process on majority class. for the process of determining SV Sets and NSV Sets by using hyperplane of negative samples. If the calculation result gives then it is categorized into SV Sets and otherwise categorized as NSV Sets.
10. SV Sets on the Minority Class will be eliminated noise and then will experience the SMOTE process to then be combined with NSV Sets on the minority class to become a new minority class.
11. NSV Sets on Majority Class will experience Multiple Random Under Sampling process and will then be combined with SV Sets on Majority Class to become the new majority class.

4 Research Framework

4.1. Proposed Taxonomy Model

Galar *et al.* presents the Ensemble Approach into a taxonomy consisting of 4 (four) groups: cost-sensitive boosting, boosting-based ensembles, bagging-based ensembles, and hybrid ensembles [3]. The complete taxonomy can be seen in Figure 4.

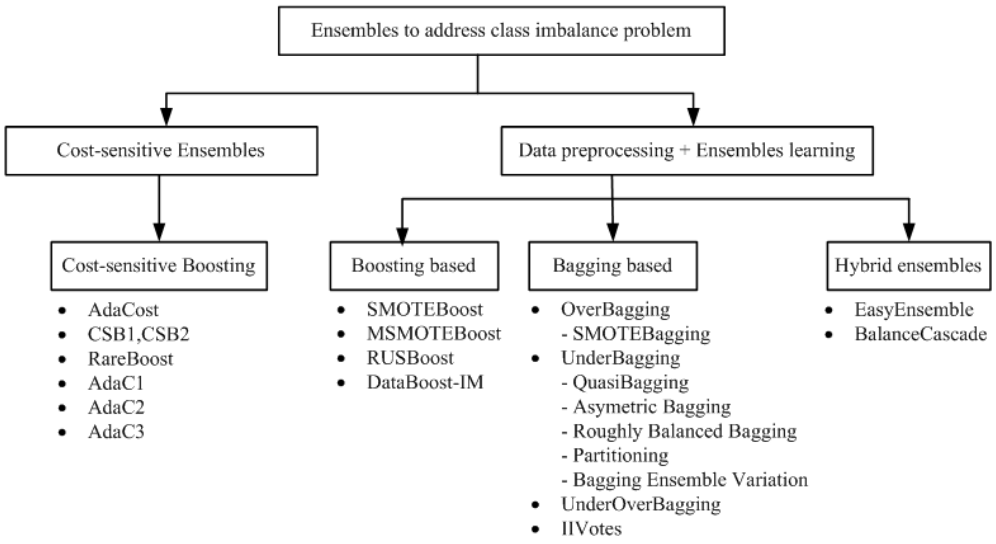

Fig. 4. Galar Taxonomy for Imbalance Learning.

The taxonomic model expressed by Galar *et al.* [3] has evolved with respect to the introduction of sampling-based ensemble methods by a number of researchers [12-17].

The new taxonomic form after considering the sampling based ensembles method expressed by [12-17] can be seen in Figure 5.

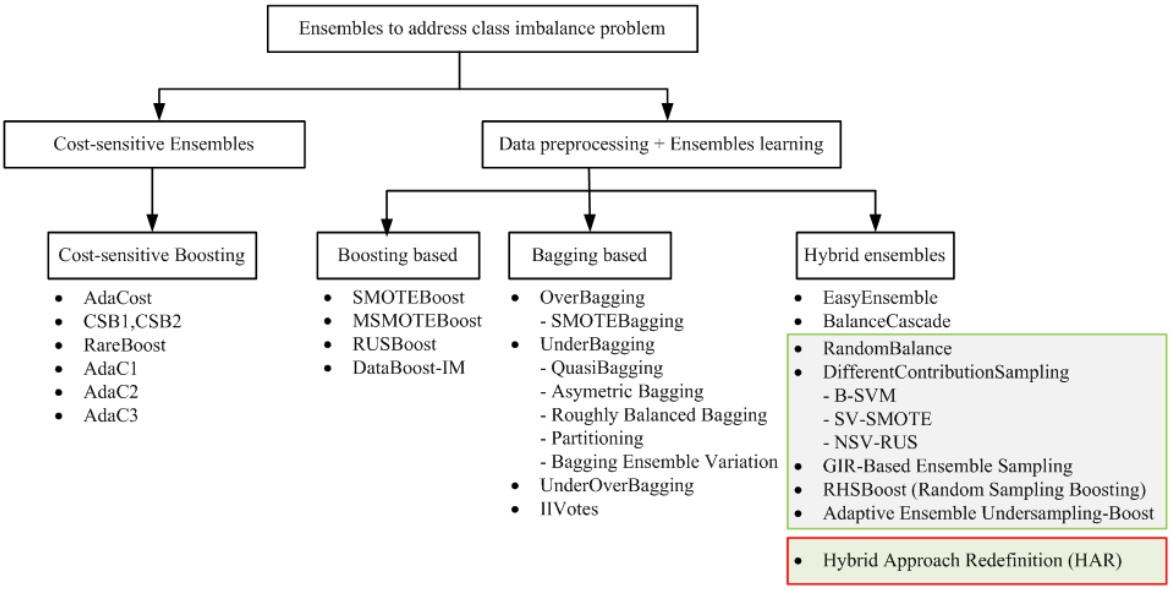

Fig. 5. Proposed Taxonomy Model.

Hartono *et al.* has used Biased Support Vector Machine and Weighted-Smote in handling class imbalance problem [18].

4.2. Hybrid Approach Redefinition (HAR) Mode

In this study, we will present the Hybrid Approach Redefinition (HAR) Model which combines Hybrid Ensembles and Sampling Based Ensembles methods. The Random Balance Ensemble Method will be used for the preprocessing stage while processing will be performed using DCS and Bagging methods.

The pseudocode of the Hybrid Approach Redefinition (HAR) Model is as follows.

Input: *S: Training Set; T:* Number of Iterations
n: Bootstrap Size
Output: Bagged Classifier: $H(x) = sign\left(\sum_{t=1}^{T} h_t(x)\right)$ where $h_t \in [-1, 1]$ are the induced classifiers
1: **for** $t = 1$ to T **do**
2: $S_t \leftarrow$ *Preprocessed Data Test* using *Random Balance Ensemble Method (n,S)*
3: Classifying S_t Using B-SVM
4: Identifying Negative Samples
5: Identifying Positive Samples
6: While (!EndofNegativeSamples) do
7: NewSVSets[]←Deleting the Noise Samples in SV Sets
8: NewNSVSets[]←Multiple Random Under-Sampling in NSV Sets
9: end while
10: For All NewSVSets and NewNSVSets do
11: New NegativeSampleSets← $NewSVSets + NewNSVSets$
12: End
13: While (!EndofPositiveSamples) do
14: SMOTESets[]←Deleting the Noise Samples in SV Sets
15: end while
16: For All SMOTESets and NewNSVSets do
17: New PositiveSampleSets← $SMOTESets + NewNSVSets$
18: End
19: For All NewNegativeSampleSets and NewPositiveSampleSets do
20: ResultDataSet← $NewNegativeSampleSets + NewPositiveSampleSets$
21: End
22: End

Hybrid Ensembles Optimization is done by combining Hybrid Ensembles and Sampling Based Ensembles methods. Where Random Balance Ensemble Method will be used for preprocessing stage while processing will be done by using Different Contribution Sampling (DCS) and Bagging method. The purpose of Hybrid Ensembles optimization is to reduce the size of the classifier and increase the diversity of data.

5 Conclusion

The conclusion of this research are as follows. First, in clustering, the class imbalance problem not only affects the accuracy of a prediction but also introduces bias in decision making process, it was indicated by decreasing the number of errors. Second, it is confirmed that the method in handling class imbalance problem must consider the size of the classifier and the data diversity. Third, This research framework aims to optimize hybrid ensembles so as to reduce classifier size and increase data diversity.

This work was supported by the Grant of Ministry of Research, Technology, and Higher Education (KEMENRISTEKDIKTI) of the Republic of Indonesia.

References

1. S.M.A. Erhahman, A. Abraham, A Review of Class Imbalance Problem, J. of Network and Innovative Computing 1, pp. 332-340 (2014)
2. Hartono, O.S. Sitompul, Tulus, and E.B. Nababan, Optimization Model of K-Means Clustering Using Artificial Neural Networks to Handle Class Imbalance Problem, *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 288, p. 012075, Jan. 2018.
3. M. Galar, A. Fernandez, E. Barrenechea & H. Bustince H, A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting, and Hybrid-Based Approachs. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews **42**, pp. 463-484 (2012)
4. Hartono, O.S. Sitompul, E.B. Nababan, Tulus, D. Abdullah, A.S. Ahmar, A New Diversity Technique for Imbalance Learning Ensembles, *Int. J. Eng. Technol.* **7** (2), pp. 478-483 (2018)
5. S. Ertekin, J. Huang, C.L. Giles, Active Learning for Class Imbalance Problem*,* Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval*,* pp. 823-824 (2007)
6. M. Alibeigi, S. Hashemi, A. Hamzeh, DBFS: An Effective Density Based Feature Selection Scheme for Small Sample Size, and High Dimensional Imbalanced Data Sets. *Data Knowledges Engineering* **81**, pp. 67-103 (2012)
7. M. Galar, A. Fernandez, E. Barrenechea, F. Herrera, EUSBoost: Enhancing Ensembles for Highly Imbalanced Data-Sets by Evolutionary Undersampling, Pattern Recognition **46**, pp. 3460-3471 (2013)
8. A. Fernandez, V. Lopez, M. Galar, M.J.D. Jesus, F. Herrera, Analyzing the Classification of Imbalanced Data-Sets with Multiple Classes: Binarization Techniques and Ad-hoc Approaches. Knowledge-Based Systems **42**, pp. 97-110 (2013)
9. M.J. Zaki, W. Meira Jr., *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press (2014)
10. S. Almuhaideb, M.E.B. Menai, Impact of Preprocessing on Medical Data Classification. Frontiers of Computer Science **10**, pp. 1082-1102 (2016)
11. B. Krawczyk, G. Schaefer, M. Wozniak, A Hybrid Cost-Senstivie Ensemble for Imbalanced Breast Thermogram Classification. Artificial Intelligence in Medicine **65**, pp. 219-227 (2015)
12. F.D.P. Jose, J.J. Rodriguez, C.G. Osorio, L.I. Kuncheva, Random Balance: Ensembles of Variable Priors Classifiers for Imbalanced Data, Knowledge-Based Systems **85**, pp. 96-111 (2015)
13. C. Jian, J. Gao, Y. Ao, A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble, Neurocomputing **193**, pp. 115-122 (2016)
14. B. Tang, H. He, GIR-Based Ensemble Sampling Approaches for Imbalanced Learning. Pattern Recognition **71**, pp. 306-319 (2017)
15. F. Ren, P. Cao, W. Li, D. Zhao, O. Zaiane*,* Ensemble Based Adaptive Over-Sampling Method for Imbalanced Data Learning in Computer Aided Detection of Microaneurysm. Computerized Medical Imaging and Graphics **55**, pp. 54-67 (2017)
16. J. Gong J, H. Kim, RHSBoost: Improving Classification Performance in Imbalance Data. Computational Statistics & Data Analysis **111**, 1-13 (2017)
17. W. Lu, S. Li, J.. Chu, Adaptive Ensemble Undersampling-Boost: A Novel Learning Framework for Imbalanced Data. Journal of Systems and Software **132**, pp. 272-282 (2017)
18. Hartono, O.S. Sitompul, Tulus, E.B. Nababan, Biased support vector machine and weighted-smote in handling class imbalance problem, International Journal of Advances in Intelligent Informatics **4***,* 1, pp. 21-27 (2018)
19. KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on)." [Online]. Available: http://sci2s.ugr.es/keel/ datasets.php. [Accessed: 29-April-2018].