# Hybrid Approach Redefinition (HAR) Method with Loss Factors in Handling Class Imbalance Problem

Hartono
Department of Computer Science
STMIK IBBI
Medan, Indonesia
hartonoibbi@gmail.com

Erianto Ongko
Department of Informatics
Akademi Teknologi Industri Immanuel
Medan, Indonesia

Opim Salim Sitompul, Tulus, Erna Budhiarti Nababan
Department of Computer Science
Universitas Sumatera Utara
Medan, Indonesia

Dahlan Abdullah
Department of Informatics
Universitas Malikussaleh
Aceh, Indonesia

***Abstract*—Class imbalance is the main problem in classification because the classification process tends to misclassify minority class which is an interesting class in another class if the training process is done to a set of instances. This problem will result in the result obtained biased towards to the class with a large number of instances. Against a number of methods proposed to overcome this class imbalance problem. One good method is the Hybrid Approach Redefinition (HAR) Method which has the advantage in overcoming the problem of class imbalance with the number of small classifiers and also the data diversity is good. This study will use the HAR Method incorporated with Loss Factors to correct the classification of most classes based on the performance evaluation of each classifier based on the F-Measure and G-Mean values. The results showed that HAR Method with Loss Factors gave better performance value compared with HAR Method without Loss Factor.**

***Keywords—class imbalance; classification; hybrid approach redefinition (HAR) method;, loss factors***

## I. INTRODUCTION

Many classifying tasks in the real world suffer from the problem of an unbalanced class in which some classes are highly represented in comparison to other classes [1]. The misclassifying of the classification process can decrease the accuracy of prediction [2]. Within the Machine Learning Research Society, the problem of learning imbalances in the classroom is considered to be one of the most important challenges [3]. Class imbalance is the main problem in classification because the classification process tends to misclassify minority class which is an interesting class in another class if the training process is done to a set of instances [4]. Class imbalance is the main problem in many applications for real-world sample recognition, In automated image monitoring, fraud detection, PC and network security, risk management, and medical diagnostics [5]. Classification ensemble has emerged as a popular learning framework to address unbalanced classification problems [6]. ensemble classifier can provide higher accuracy and durability than a single classifier by combining different classifier [7].

Boosting is a common static ensemble learning algorithm initiated to effectively promote a weak learner who behaves a little better than random guessing in a stronger ensemble [5]. SMOTE is also involved in and/or expanded to an ensemble-based method [6]. In Boosting, there is a weight in the process of learning so that the base classifiers in the file are aimed at correctly classifying more important samples as iterative events continue. Samples that are incorrectly classified in each iteration become more important for further iteration, and a more precise primary classifier acquires a higher share in the final decision [5]. Soleymani *et al.* [4] proposed Loss Factors to improve the performance of Boosting algorithms in unbalanced data. The modified loss factor is integrated into the weight update formula and determines the contribution of the classifier in the final class prediction to avoid the distortion of the majority class.

Hybrid ensemble is a method that combines both bagging and boosting [8]. Hybrid Approach Redefinition (HAR) method which is basically a hybrid ensemble and argues that this method specially designed to increase diversity and have an impact to the performance of imbalance learning and HAR Method also use SMOTEBoost in their process [9] without loss factors for boosting. This study will use the HAR Method incorporated with Loss Factors to correct the classification of most classes based on the performance evaluation of each classifier based on the F-Measure and G-Mean values [8].

## II. RELATED WORKS

Another problem in Boosting-Based Ensembles is that they may suffer the bias towards a negative class because the loss factor that governs their learning process is gained based on the weighted precision. In cases of imbalance, the weighted accuracy reflects the ability to correctly classify negative samples more than positive [5]. One of the solutions is using

978-1-5386-5280-0/18/$31.00 ©2018 IEEE

Cost-Sensitive Approach that defines the misclassification costs for different classes[10]. Prusty *et al*. [11] has proposed a Weighted SMOTE where oversampling of each minority data sample is carried out based on the weight assigned to it. Biased Support Vector Machine and Weighted SMOTE had been used in handling class imbalance problem and give better accuracy and sensitive compared to genera Support Vector Machine [12]. In contrast, Cost-Free techniques modify learning algorithms by increasing factor loss calculation without taking into account cost factors [13]. Another Cost-Free Approach [4] modifies the calculation of the loss factor of the Boosting algorithm by the F-measure, the most commonly used performance evaluation measures for imbalance learning.

## III. METHODOLOGY

The data used in this research are Ecoli0vs1, Glass0, and NewThyroid2 from KEEL-Dataset Repository and UCI Machine Learning Repository. The dataset used is a dataset with varying degrees of imbalanced (imbalanced ratios). The Research Method can be seen in Fig. 1

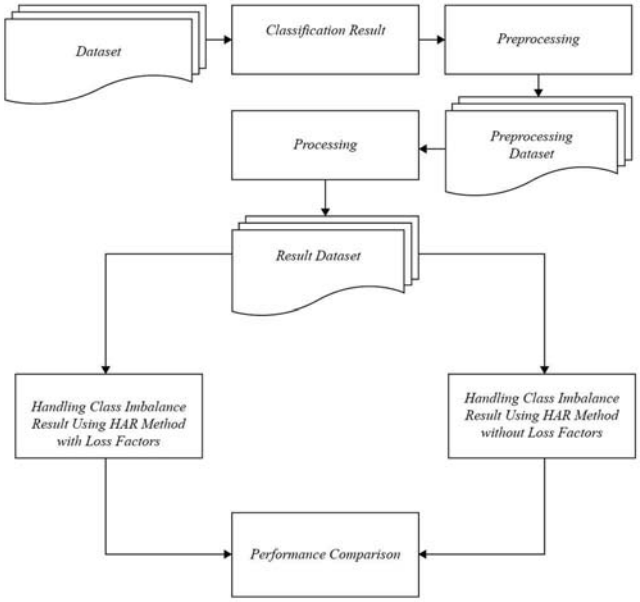


Fig. 1. Research Method

In Fig. 1 it can be seen that the process begins with the selection of the dataset. The existing dataset will undergo a classification process with one of the classification algorithms such as K-Means. The result of classification containing the class imbalance will undergo preprocessing step by using Random Balance Ensemble Method[14]. The result of preprocessing in the form of preprocessing dataset will undergo processing stages by using Different Contribution Sampling[15]. The results of this processing will be compared with the HAR Method, to see if the HAR Method with Loss Factors is better than the HAR Method without Loss Factors.

### *A. Hybrid Approach Redefinition (HAR) Method*

Hybrid Approach Redefinition (HAR) will use the Random Balance Ensemble Method for preprocessing steps in order to maintain data diversity and preprocessing will be done using UnderBagging and Different Contribution Sampling (DCS) methods in order to reduce the size of the classifier [9]. The process of Hybrid Approach Redefinition (HAR) Method can be seen if Fig. 2, Fig. 3, and Fig. 4 [9].


Fig. 2. Preprocessing Stage at HAR Method

In Fig. 2, Fig. 3, and Fig. 4 the steps shown in the HAR Method are divided into 3 sections: Preprocessing Stage, Processing Stage, and Evaluation Stage. In the Preprocessing Stage stage, the Random Balance Ensemble Method is used which combines Random Under Sampling with SMOTE Boost[14]. This preprocessing stage begins with determining the size of the majority and minority classes, and then a new minority class will be generated by the Random Undersampling method. In the Random Under Sampling method if the new majority class size is smaller than the old majority class size, take some samples from majority class and move to minority class and if otherwise take samples from minority class and move to majority class. This preprocessing stage is expected to reduce the number of classifier.

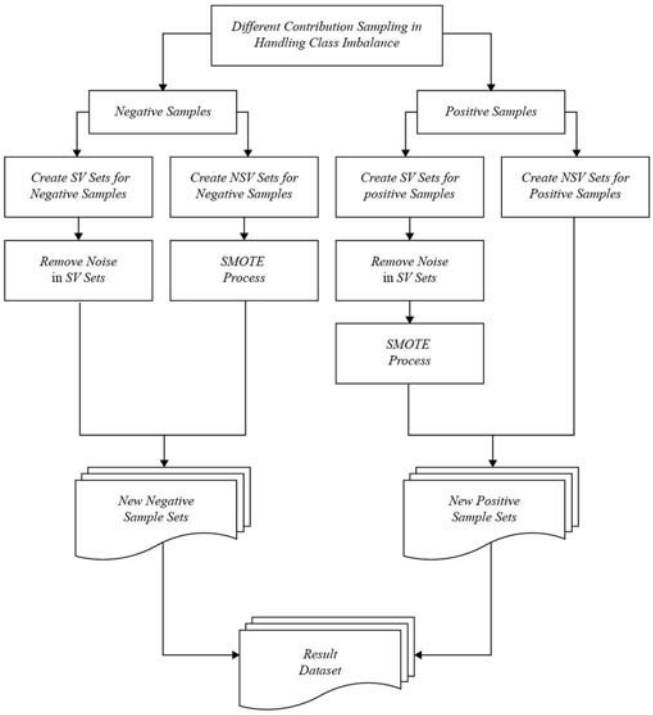


Fig. 3. Processing Stage at HAR Method

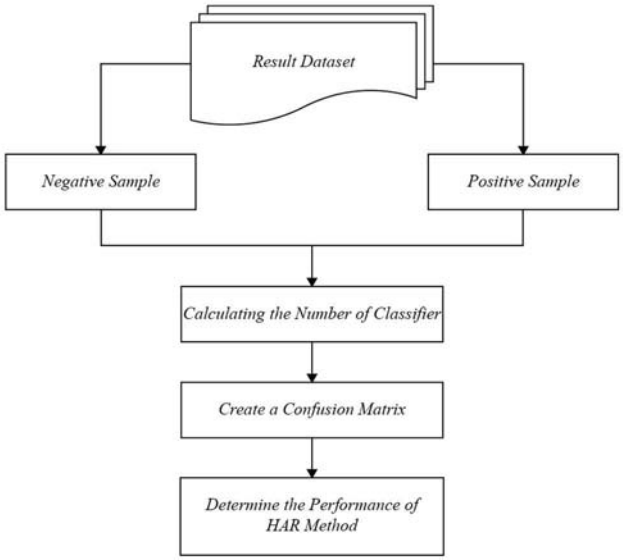


Fig. 4. Evaluation Stage at HAR Method

The whole process of forming the majority and minority of the new class is done using SMOTEBoost. In Fig. 3, the processing is done using Different Contribution Sampling (DCS) method[15]. In the DCS method, either Negative Samples or Majority Class or Positive Samples or Minority Class will be separated into SV Sets and NSV Sets. NSV Sets on Majority Class will undergo SMOTE process and the same will be done for SV Sets on Minority Class. Processing stages involving SMOTEBoost with Loss Factors for Boosting are expected to achieve good data diversity. The results of the processing will result in the Result Dataset. Result The dataset obtained then will be measured by its performance.

### *B. Loss Factors for Boosting*

Another cost-free approach [4] modifies the calculation of the loss factor of the Boosting algorithm using F-Measure, the measures most frequently used for the evaluation of performance in the learning of class imbalance. This loss factor can be integrated into any Boosting ensemble technique.

To aim this, we separated the weight vector into two vectors for minority and majority class as shown in (1) and (2) respectively [4].

$$\boldsymbol{W}_t^+ = \{w_t(i), i = 1,\dots,N | y_i = 1\}, \quad (1)$$
$$\boldsymbol{w}_t^- = \{w_t(i), i = 1,\dots,N | y_i = -1\} \quad (2)$$

Then calculated the True Positive, False Positive, True Negative, and False Negative for weighted versions as shown in (3) – (6).

$$TP_t = \sum_{(j,Y_j):Y_j=1} \boldsymbol{w}_t^+(j), j = 1,\dots,N^+ \quad (3)$$
$$FP_t = \sum_{(j,Y_j):Y_j=1} \boldsymbol{w}_t^-(j), j = 1,\dots,N^- \quad (4)$$
$$TN_t = \sum_{(j,Y_j):Y_j=-1} \boldsymbol{w}_t^-(j), j = 1,\dots,N^- \quad (5)$$
$$FN_t = \sum_{(j,Y_j):Y_j=-1} \boldsymbol{w}_t^+(j), j = 1,\dots,N^+ \quad (6)$$

And then, the accuracy of classifier is calculated in terms of $F_\beta - Measure$ as shown in (7).

$$A_F = \frac{(1+\beta^2)TP_t}{(1+\beta^2)TP_t + FP_t + \beta^2 FN_t} \quad (7)$$

To measure the error of classifiers, the corresponding loss factor are defined respectively as shown in (8).

$$L_t = 1 - A_F = \frac{FP_t + \beta^2 FN_t}{(1+\beta^2)TP_t + FP_t + \beta^2 FN_t)} \quad (8)$$

Determination of weight vector into two vectors for minority and majority class is expected to obtain better diversity data because it can reduce the number of misclassification which is only centered on one class.

### *C. SMOTEBoost with Loss Factor*

The Pseudocode of the SMOTE Algorithm with Loss Factor is as follows [4].

**Input**: Training Set S={($x_i$, $y_i$); $i$ = 1,...,$N$}
$N = N^+ + N^-$
Number of Iteration $\boldsymbol{T}$
Number of SMOTE $N$
Number of Nearest Neighbors $K$

**Process**:
1: if N < 100
2: then Randomize the T minority class samples
3: $T$ = (N/100) * $T$
4: N = 100

5: end if
6: N = (int)(N/100)
7: k = Number of nearest neighbors
8: numattrs = Number of attributes
9: Sample[ ][ ]: Minority Class Sample
10: newindex: 0
11: Synthetic[ ][ ]: array for synthetic samples
12: for i ← 1 to $S_P$
13: Compute k nearest neighbors
14: Populate(N, i, nnarray)
15: end for
16: $\mathbf{W}_1(i) = 1/N$
17: for t ← $T$ *do*
18: Create new training set $S_t'$ with weight distribution $W_t'$ using RS
19: Train classifier $C_t$ on $S_t'$ with $W_t'$
20: Test $C_t$ on **S** and produce labels $\{Y_i, i = 1,..., N\}$
21: Define weight vectors $W_t^+$ and $W_t^-$ as in Eqs. (1) and (2)
22: Compute loss factor $L_t$ with Eqs. (8)
23: If $L_t > \frac{N^-}{(1+\beta^2)N^+ + N^-}$ go to Step (18)
23: Calculate the weight update parameter: $\alpha_t = \frac{L_t}{1-L_t}$
24: Update $W_{t+1}(i) = W_t(i)\alpha_t^{\frac{1}{2}|y_i - Y_i|}$
25: Normalize $W_{t+1}$ such that: $\sum W_{t+1} = 1$

*D. HAR Method with Loss Factors*

The Pseudocode of the HAR Method with Loss Factor is as follows [9].

**Preprocessing**
Using *Random Undersampling* + SMOTEBoost (*Random Balance Ensemble Method*) with Loss Factor
*Pseudocode*
**Input**: Total Size *totalSize*
Number of Majority $S_N$
Number of Minority $S_P$
**Process**:
*totalSize* ← |S|
$S_N$ ← {($x_i$, $y_i$) ε S|$y_i$ = -1}
$S_P$ ← {($x_i$, $y_i$) ε S|$y_i$ = +1}
*majorityClass* ← |$S_N$|
*minorityClass* ← |$S_P$|
*newMajorityClass* ← Random integer between 2 and *totalSize*-2
*newMinorityClass* ← *totalSize* – *newMajorityClass*
**if** *newMajorityClass* <*majorityClass* **then**
$S'$ ← $S_P$
Take a random sample of size *newMajorityClass* from $S_N$, add the sample to $S'$
Create *newMinorityClass* – *minorityClass* artificial using **SMOTEBoost with Loss Factor**
**else**
$S'$ ← $S_N$
Take a random sample of size *newMinorityClass* from $S_P$, add the sample to $S'$
create *newMajorityClass* – *majorityClass* artificial using **SMOTEBoost with Loss Factor**

**end if**
**return** $S'$

**Processing**
Using UnderBagging dan Different Contribution Sampling
**Input**: *S: Training Set;*
*T:* Number of Iterations
*n*: Bootstrap Size
**Output:** Bagged Classifier: $H(x)$ =
*sign* $(\sum_{t=1}^{T} h_t(x))$ where $h_t$ [-1, 1] are the induced classifiers
**Process:**
**for** *t* = 1 to T **do**
$S_t$ *Preprocessed Data Test* using *Random Balance Ensemble Method (n,S)*
Classifying $S_t$ Using B-SVM
Identifying Negative Samples
Identifying Positive Samples
While (!EndofNegativeSamples) do
NewSVSets[]Deleting the Noise Samples in SV Sets
NewNSVSets[]Multiple Random Under-Sampling in NSV Sets
end while
For All NewSVSets and NewNSVSets do
New NegativeSampleSets
End For
While (!EndofPositiveSamples) do
SMOTESets[]Deleting the Noise Samples in SV Sets
end while
For All SMOTESets and NewNSVSets do
New PositiveSampleSets
End For
For All NewNegativeSampleSets and New PositiveSampleSets do
ResultDataSet
End For
End For

The main difference from the HAR Method with Loss Factors and without Loss Factors is the determination of weight vector into two vectors for minority and majority class in the HAR Method with Loss Factors which is expected to obtain better diversity data because it can reduce the number of misclassifications that are only focused on one class. If the number of misclassifications concentrated only on one class will certainly gain low data diversity, and Loss Factors for boosting can minimize it by determining the right weight for each minority and majority class.

*E. Measurement of Performance*

Measurement of Performance is using F-Measure, G-Means, and Q-Statistics as shown in (9) - (17) [10] and [16].

$$True\ Negative\ Rate\ (TNrate) = \frac{TN}{TN + FP} \quad (9)$$

$$False\ Negative\ Rate\ (F\ Nrate) = \frac{FN}{TP+} \quad (10)$$

$$Positive\ Predictive\ Value\ (PPValue) = \frac{TP}{TP + FP} \quad (11)$$

$$Negative\ Predictive\ Value\ (NPValue) = \frac{TN}{TN + FN} \quad (12)$$

$$Recall = TPrate = \frac{TP}{TP + FN} \quad (13)$$

$$Precision = PPValue = \frac{TP}{TP + FP} \quad (14)$$

$$F\text{-}Measure = \frac{2RP}{R+P} \quad (15)$$

$$G\text{-}Mean = \sqrt{TPrate\ .\ TNrate} \quad (16)$$

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \quad (17)$$

## IV. Experimental Process

### A. Dataset Description

The data used in this research are Ecoli0vs1, Glass0, and New-Thyroid2. The description about dataset can be seen in Table I [17].

TABLE I. Dataset Description

| *Dataset* | #Ex | #Atts | (%Min;%Max) | IR |
|---|---|---|---|---|
| Ecoli0vs1 | 220 | 7 | (35.00,65.00) | 1.86 |
| Glass0 | 214 | 9 | (32.71,67.29) | 2.06 |
| New-Thyroid2 | 215 | 5 | (16.89,83.11) | 4.92 |

### B. Testing

The test results of F-Measure, G-Means, and Q-Statistics for Class Imbalance Handling in Clustering Result of Ecoli0vs1 Dataset can be seen in Table II.

TABLE II. Testing Result of Ecoli0vs1 Dataset

| Testing Number | *HAR Method* | | | HAR Method with Loss Factor | | |
|---|---|---|---|---|---|---|
| | *F-Measure* | *G-Means* | *Q-Statistics* | *F-Measure* | *G-Means* | *Q-Statistics* |
| 1 | 0.73 | 0.79 | 0.05 | 0.83 | 0.78 | 0.15 |
| 2 | 0.71 | 0.74 | 0.97 | 0.81 | 0.76 | 0.89 |
| 3 | 0.77 | 0.80 | 0.25 | 0.75 | 0.77 | 0.16 |
| 4 | 0.89 | 0.90 | 0.14 | 0.87 | 0.82 | 0.21 |
| 5 | 0.85 | 0.88 | 0.33 | 0.97 | 0.83 | 0.15 |
| 6 | 0.84 | 0.86 | 0.01 | 0.77 | 0.74 | 0.38 |
| 7 | 0.73 | 0.77 | 0.25 | 0.87 | 0.85 | 0.21 |
| 8 | 0.67 | 0.65 | 0.19 | 0.76 | 0.74 | 0.22 |
| 9 | 0.80 | 0.82 | 0.31 | 0.74 | 0.77 | 0.27 |
| 10 | 0.89 | 0.90 | 0.14 | 0.76 | 0.81 | 0.17 |
| Average | 0.788 | 0.811 | 0.264 | 0.813 | 0.814 | 0.252 |

The test results of F-Measure, G-Means, and Q-Statistics for Class Imbalance Handling in Clustering Result of Glass0 Dataset can be seen in Table III. The test results of F-Measure, G-Means, and Q-Statistics for Class Imbalance Handling in Clustering Result of New-Thyroid2 Dataset can be seen in Table IV.

TABLE III. Testing Result of Glass0 Dataset

| Testing Number | *HAR Method* | | | HAR Method with Loss Factor | | |
|---|---|---|---|---|---|---|
| | *F-Measure* | *G-Means* | *Q-Statistics* | *F-Measure* | *G-Means* | *Q-Statistics* |
| 1 | 0.67 | 0.69 | 0.42 | 0.73 | 0.71 | 0.25 |
| 2 | 0.85 | 0.36 | 0.65 | 0.82 | 0.32 | 0.49 |
| 3 | 0.67 | 0.6 | 0.24 | 0.69 | 0.55 | 0.37 |
| 4 | 0.68 | 0.68 | 0.44 | 0.81 | 0.71 | 0.44 |
| 5 | 0.7 | 0.7 | 0.3 | 0.65 | 0.76 | 0.34 |
| 6 | 0.69 | 0.7 | 0.45 | 0.62 | 0.67 | 0.51 |
| 7 | 0.69 | 0.63 | 0.51 | 0.8 | 0.68 | 0.47 |
| 8 | 0.9 | 0.81 | 0.11 | 0.79 | 0.82 | 0.22 |
| 9 | 0.72 | 0.73 | 0.4 | 0.73 | 0.76 | 0.44 |
| 10 | 0.67 | 0.56 | 0.46 | 0.77 | 0.65 | 0.4 |
| Average | 0.724 | 0.646 | 0.398 | 0.741 | 0.663 | 0.393 |

TABLE IV. Testing Result of New-Thyroid2 Dataset

| Testing Number | *HAR Method* | | | HAR Method with Loss Factor | | |
|---|---|---|---|---|---|---|
| | *F-Measure* | *G-Means* | *Q-Statistics* | *F-Measure* | *G-Means* | *Q-Statistics* |
| 1 | 0.74 | 0.87 | 0.33 | 0.75 | 0.85 | 0.25 |
| 2 | 0.8 | 0.81 | 0.94 | 0.84 | 0.69 | 0.49 |
| 3 | 0.65 | 0.77 | 0.36 | 0.65 | 0.84 | 0.37 |
| 4 | 0.7 | 0.83 | 0.76 | 0.77 | 0.77 | 0.44 |
| 5 | 0.67 | 0.79 | 0.44 | 0.73 | 0.85 | 0.34 |
| 6 | 0.8 | 0.81 | 0.57 | 0.79 | 0.87 | 0.51 |
| 7 | 0.86 | 0.88 | 0.23 | 0.82 | 0.85 | 0.47 |
| 8 | 0.65 | 0.8 | 0.15 | 0.69 | 0.88 | 0.22 |
| 9 | 0.79 | 0.81 | 0.94 | 0.82 | 0.82 | 0.44 |
| 10 | 0.74 | 0.84 | 0.57 | 0.88 | 0.85 | 0.4 |
| Average | 0.74 | 0.821 | 0.529 | 0.774 | 0.827 | 0.393 |

## V. Result and Discussion

Based on the research results, it can be seen that the HAR Method with Loss Factor gives slightly better results than the HAR Method. This can be seen through the value of F-Measure, G-Means, and Q-Statistics obtained through HAR Method with Loss Factor better than HAR Method. In general, F-Measure value, G-Means shows better classifier performance, while lower Q-Statistics shows better data diversity.

Through the determination of the weight vector into two vectors for minority and majority class on HAR Method with Loss Factors expected to obtain better data diversity because it will be able to reduce the number of misclassification which only concentrated on one class. Better results on the HAR Method with Loss Factor show that Loss Factor has been able to correct the classification process on the Majority Class and the Minority Class.

## VI. Conclusion

The conclusion of this research is as follows. First, On the handling of the class imbalance need to pay attention to the classification process in the majority class especially with attention to Fb-Measure or G-Mean. Second, Loss Factor Calculations that pay attention to Fb-Measure can improve performance in the class imbalance handling process shown by F-Measure and G-Means after better class imbalance handling. Third, the results also show that the HAR Method with Loss

Factor can obtain better Q-Statistics value which means that the data diversity obtained is good.

## ACKNOWLEDGMENT

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

## REFERENCES

[1] S. García, Z.-L. Zhang, A. Altalhi, S. Alshomrani, and F. Herrera, "Dynamic ensemble selection for multi-class imbalanced datasets," *Inf. Sci.*, vol. 445–446, pp. 22–37, Jun. 2018.

[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] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[4] R. Soleymani, E. Granger, and G. Fumera, "Loss factors for learning Boosting ensembles from imbalanced data," 2016, pp. 204–209.

[5] R. Soleymani, E. Granger, and G. Fumera, "Progressive boosting for class imbalance and its application to face re-identification," *Expert Syst. Appl.*, vol. 101, pp. 271–291, Jul. 2018.

[6] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.

[7] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, Feb. 2010.

[8] 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 Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[9] Hartono, D. Abdullah, and A. S. Ahmar, "A New Diversity Technique for Imbalance Learning Ensembles," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 478–483, Apr. 2018.

[10] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, Dec. 2007.

[11] M. R. Prusty, T. Jayanthi, and K. Velusamy, "Weighted-SMOTE: A Modification to SMOTE for Event Classification in Sodium Cooled Fast Reactors," *Prog. Nucl. Energy*, vol. 100, pp. 355–364, 2017.

[12] H. Hartono, O. S. Sitompul, T. Tulus, and E. B. Nababan, "Biased support vector machine and weighted-smote in handling class imbalance problem," *Int. J. Adv. Intell. Inform.*, vol. 4, no. 1, pp. 21–27, Apr. 2018.

[13] M.-J. Kim, D.-K. Kang, and H. B. Kim, "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1074–1082, Feb. 2015.

[14] J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, "Random Balance: Ensembles of variable priors classifiers for imbalanced data," *Knowl.-Based Syst.*, vol. 85, pp. 96–111, Sep. 2015.

[15] C. Jian, J. Gao, and Y. Ao, "A new sampling method for classifying imbalanced data based on support vector machine ensemble," *Neurocomputing*, vol. 193, pp. 115–122, Jun. 2016.

[16] G. U. Yule, "On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society, &c," *Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character*, vol. 194, pp. 257–319, 1900.

[17] "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: 22-May-2018].