International Journal of Engineering & Technology, 7 (2) (2018) 478-483

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



A New Diversity Technique for Imbalance Learning Ensembles

Hartono[1,2]*, Opim Salim Sitompul[2], Erna Budhiarti Nababan[2], Tulus[3], Dahlan Abdullah[4], Ansari Saleh Ahmar[5]

[1]*Department of Computer Science, STMIK IBBI, Medan, Indonesia*
[2]*Department of Computer Science, Universitas Sumatera Utara, Medan, Indonesia*
[3]*Department of Mathematics, Universitas Sumatera Utara, Medan, Indonesia*
[4]*Department of Informatics, Universitas Malikussaleh, Aceh, Indonesia*
[5]*Department of Statistics, Universitas Negeri Makassar, Makassar, Indonesia*
**Corresponding author E-mail:hartonoibbi@gmail,com*

This Paper is based on a presentation given by the authors at "Workshop of KO2PI" held from 19 January 2018 to 20 January 2018 in Indonesia.

Abstract

Data mining and machine learning techniques designed to solve classification problems require balanced class distribution. However, in reality sometimes the classification of datasets indicates the existence of a class represented by a large number of instances whereas there are classes with far fewer instances. This problem is known as the class imbalance problem. Classifier Ensembles is a method often used in overcoming class imbalance problems. Data Diversity is one of the cornerstones of ensembles. An ideal ensemble system should have accurrate individual classifiers and if there is an error it is expected to occur on different objects or instances. This research will present the results of overview and experimental study using Hybrid Approach Redefinition (HAR) Method in handling class imbalance and at the same time expected to get better data diversity. This research will be conducted using 6 datasets with different imbalanced ratios and will be compared with SMOTEBoost which is one of the Re-Weighting method which is often used in handling class imbalance. This study shows that the data diversity is related to performance in the imbalance learning ensembles and the proposed methods can obtain better data diversity.

Keywords: *Class Imbalance, Classifier Ensembles. Data Diversity, Hybrid Approach Redefinition*

1. Introduction

In the classification, the dataset is said to be imbalanced when there is a class with a smaller amount of data than the other [1]. Class with larger amount of data is named as majority class whereas class with small amount of data is named as minority class. The problem of class imbalance in the classification process has been a challenge in the classification process and attracted the attention of a number of researchers [2]. Imbalanced data classification is difficult because standard classifier is driven by accuracy, where minority classes are often ignored [3] and traditional classifiers generally favor the majority class which has a large number of instances [4]. In clustering, this problem not only affects the accuracy of a prediction but also introduces bias in decision-making process [5].

in [6], the approaches to dealing with class imbalance problems can be divided into four categories: Algorithm-Level, Data-Level, Cost-Sensitive, and Classifier Ensembles. The main idea of ensemble learning is to train the pool of base classifiers on different versions of training datasets and aggregate their decisions to classify unknown instances. Ensemble methods for imbalanced learning tackle the imbalance problem using techniques like re-weighting, Oversampling, and Undersampling. As one of the most popular ensemble approaches, boosting [7] re-samples adaptively the examples according to their weights, and produces a highly accurate ensemble of classifiers whose individual classifier holds a moderate accuracy and this method often called as Re-Weighting. Diversity is one of the cornerstones of ensembles [3].

An ideal ensemble system should have accurrate individual classifiers and if there is an error it is expected to occur on different objects or instances [3]. In this paper we propose Hybrid Approach Redefinition (HAR) method which is basically a hybrid ensembles and argue that this method especially designed to increase diversity and have an impact to the performance of imbalance learning. This research will be conducted using 6 datasets with different imbalanced ratios and will be compared with SMOTEBoost which is one of the Re-Weighting method which is often used in handling class imbalance.

The rest of this paper is organized as follows. In Section 2 we will provide related works in increasing diversity. In Section 3 we describe the methodology used in this research and in Section 4 we provide the experimental process performed in this research. Results and discussion are given in Section 5 and finally, we conclude the research in Section 6.

2. Related Works

Ensemble of Classifiers is the process of combining multiple classifiers, termed as base classifiers. The purpose of the ensembles is to get better performance than using single classifiers [8]. In the process of generating good ensembles, it is important to not only generate good base classifiers, but also the resulting classifier must be diverse, this means that for the same instance, the base

Copyright © 2016 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

classifier may generate errors in the form of misclassification and the error should be in different instances. Diversity is essential in order to build an accurate ensemble of classifiers [3].
[9] and [10] using the method of F-Measure, G-Means, and Q-Statistic for the determination of data diversity. If only performance on positive samples is taken into account then F-Measure measurement is used and if the diversity problem will pay attention to positive samples or negative samples then use G-Mean [10]. F-Measure and G-Mean are used to describe performance trends in different degrees of diversity. Whereas Q-Statistics tends to be used for diversity measurement because its form is easy to understand [11].

3. Methodology

The data used in this research are Iris, Balanced Scale Weight & Distance Database, Haberman, Thyroid, New-Thyroid1, and New-Thyroid2 from KEEL-Dataset Repository and UCI Machine Learning Repository. This research will be done by Hybrid Approach Redefinition (HAR) method. In general, the methodology consists of four stages: selection and preparation of dataset, preprocessing, processing, and testing and evaluation.
In the selection and preparation of dataset will be determined a number of datasets to be used. The dataset used is a dataset with varying degrees of imbalanced (imbalanced ratios). Instance on the dataset will then undergo clustering process with one of clustering algorithm that is K-Means Clustering. If the clustering results indicate a problem of class imbalance then the process will proceed with the handling of the class imbalance. In the preprocessing stage will be done by using Random Balance Ensemble Method. The method of Random Balance Ensemble Method is done by using Random Under Sampling method and also SMOTEBoost. The result of this pre-processing step is a preprocessing dataset which will then proceed to the processing stage. In the processing stage will be done by using Different Contribution Sampling (DCS) method. In this DCS method the number of classifier has been reduced because it has undergone the preprocessing stage. This DCS method divides the instances of the Minority Class and Majority Class into 2 (two) sections: Support Vector Sets (SV Sets) and Non Support Vector Sets (NSV Sets). Where NSV Sets in Majority Class will be processed Random Under Sampling and SV Sets on Minority Class will be done SMOTEBoost process so that the data diversity will be good enough. After undergoing Processing step will produce result dataset. Based on the post clustering result dataset using Hybrid Approach Redefinition (HAR) it will be tested to compare post clustering results by using SMOTEBoost especially in Data Diversity.
The general architecture of the proposed method used is depicted in Fig. 1

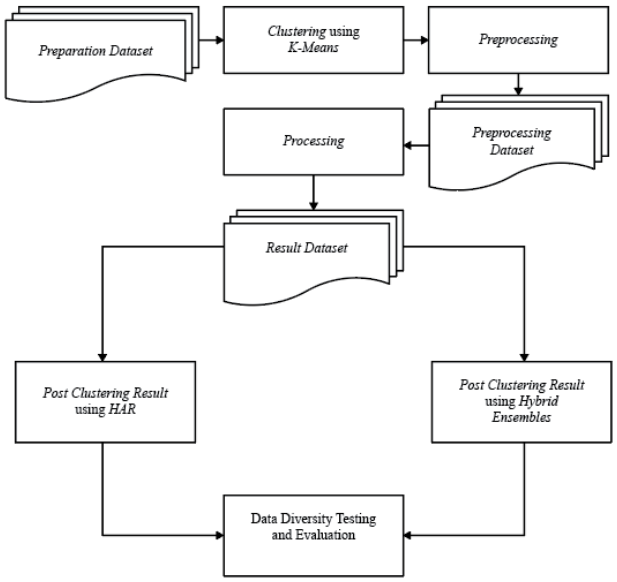


Fig. 1: The General Architecture

3.1. Hybrid Approach Redefinition (HAR) Method

The process in the selection and preparation of dataset and preprocessing stage can be seen in Fig. 2.

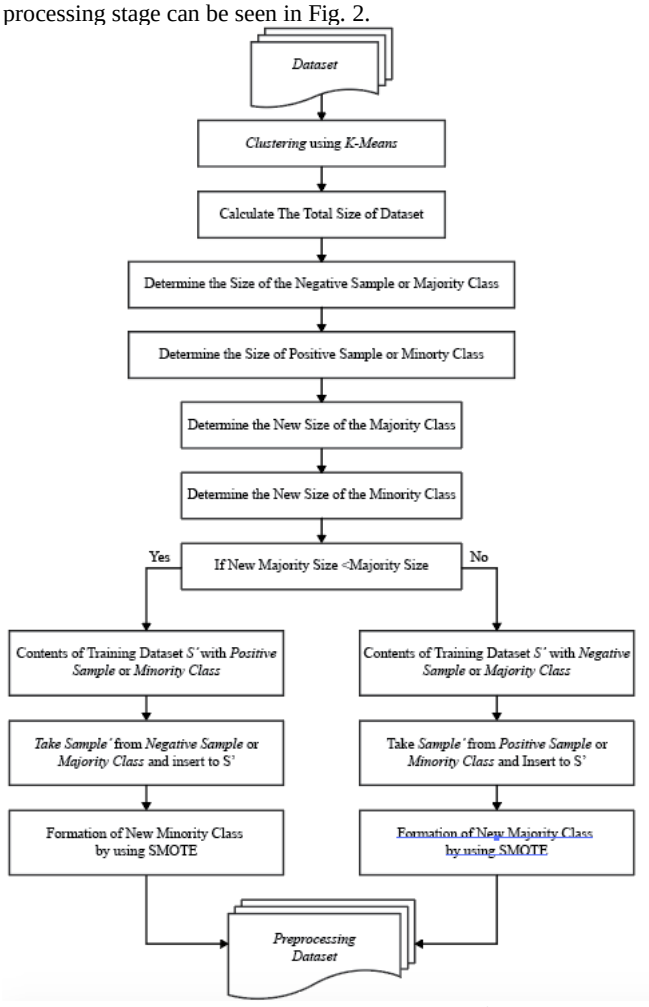


Fig. 2: The Process in The Selection and Preparation of Dataset and Preprocessing Stage in HAR Method

In Fig. 2. it can be seen that the selected dataset will experience clustering with K-Means. Clustering results that contain problems

of imbalance class will experience the process of handling class imbalance problem with Hybrid Approach Redefinition (HAR) starting from Pre-processing stage. In the pre-processing stage, the Random Balance Ensemble Method using Random Under Sampling and SMOTEBoost is used. In this pre-processing stage, the first thing to do is to obtain the number of members of the Majority Class or Negative Samples stored in the SN variable as well as the number of members of the Minority Class or Positive Samples stored in the SP variable.

After that determined the size of the new Majority Class (Negative Samples) by generating random numbers from 2 to TotalSize-2. The size of the new Minority Class (Positive Samples) is equal to TotalSize minus the size of the new Majority Class.

If the size of the new Majority Class is smaller than the size of the old Majority Class then it means the size of the Majority Class is still larger than the Minority Class it will be established New Minority Class by filling the training dataset S' with Minority Class and then do the sampling process with Random Under Sampling to retrieve a number of samples from the Majority Class and then based on the S 'training dataset the Instance on Majority Class will undergo the SMOTE process and retrieve some data from the Majority Class to be entered into the S' training dataset based on data with the nearest proximity level to the Minority Class. Based on the results of the SMOTE process, a pre-processing dataset will be generated.

Conversely, if the size of the new Majority Class is larger than the size of the old Majority Class then it means that now the position is Minority Class has a larger amount of data than the Minority Class. So the process is to fill the train dataset S 'with Majority Class, then will do the sampling process to Minority Class by using Random Under Sampling to take some samples from Minority Class and then according on training dataset S' then the instance of minority class will undergo the SMOTE process and then some data with the nearest proximity level to the Majority Class will be incorporated into the Majority Class. Based on the results of the SMOTE process, a pre-processing dataset will be generated.

The process in the processing stage can be seen in Fig. 3.

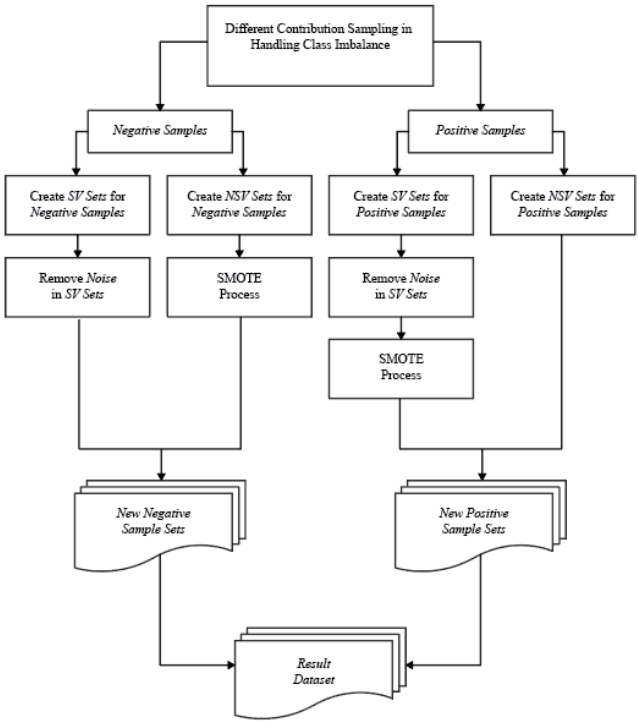


Fig. 3: The Process in Processing Stage in HAR Method

In Fig. 3. it can be seen that, once the pre-processing dataset has been generated, the pre-processing of the dataset will enter the processing stage. In the processing stages, the instance of the dataset in both the minority class and the majority class will be grouped into Support Vector (SV) Sets and Non Support Vector (NSV) Sets. Then SV Sets on Majority Class or Negative Samples will experience noise removal process while NSV Sets in Majority Class will undergo Multiple Random Under Sampling process, the instance in dataset which gives the biggest NSV value based on Multiple Random Under Sampling process will be inserted into NSV Sets on Minority Class.

The same is done in the Minority Class, where SV Sets on the Minority Class will experience the process of noise removal, then SV Sets in the Minority Class will undergo SMOTE process and instances in SV Sets with the lowest proximity will be entered into the Majority Class.

The proposed Hybrid Approach Redefiniton (HAR) Method algorithm is as follows.

Preprocessing using Random Balance Ensemble Method

The algorithm of Random Balance Ensemble Method [12]

Require: Set S of examples $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_1 \varepsilon X$ and $y_i \varepsilon Y = \{-1, +1\}$ (+1: positive or minority class, -1: negative or majority class), neighbours used in SMOTE, *k*

Ensure: New set *S*′ of examples with Random Balance

1: *totalSize* ← |S|
2: $S_N \leftarrow \{(x_i, y_i) \varepsilon S | y_i = -1\}$
3: $S_P \leftarrow \{(x_i, y_i) \varepsilon S | y_i = +1\}$
4: *majoritySize* ← $|S_N|$
5: *minoritySize* ← $|S_P|$
6: *newMajoritySize* ← Random integer between 2 and *totalSize*-2
7: *newMinoritySize* ← *totalSize* – *newMajoritySize*
8: **if** *newMajoritySize* <*majoritySize* **then**
9: $S' \leftarrow S_P$
10: Take a random sample of size *newMajoritySize* from S_N, add the sample to S'
11: Create *newMinoritySize* – *minoritySize* artificial using SMOTE
12: **else**
13: $S' \leftarrow S_N$
14: Take a random sample of size *newMinoritySize* from S_P, add the sample to S'
15: create *newMajoritySize* – *majoritySize* artificial using SMOTE
16: **end if**
17: **return** S'

Processing using Different Contribution Sampling

The algorithm of Different Contribution Sampling [7]

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

```
26:     For All SMOTESets and NewNSVSets do
27:          New PositiveSampleSets
28:     End For
29:     For All NewNegativeSampleSets and New
30:     PositiveSampleSets do
31:          ResultDataSet
32:     End For
33: End For
```

3.2. SMOTEBoost

The SMOTEBoost algorithm is as follows [13].

Input: Number of Minority S_P
Number of SMOTE N

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 SP
13:    Compute k nearest neighbors
14:    Populate(N, i, nnarray)
15: end for
16:  while N ≠ 0 do
17:     Choose a random number between 1 and k, call it nn
18:     for attr ← 1 to numattrs
19:        if attr == Continuous feature
20:              dif = Sample[nnarray[nn]][attr] - Sample[i][attr]
21:              gap = random number between 0 & 1
22:              Synthetic[newindex][attr] = Sample[i][attr] + gap *
dif
23:         else
24:              attr_value = majority vote for the attr values be-
tween i and nn.
25:              If no majority then choose at random.
26:              Synthetic[newindex][attr] = attr_value
27:     end for
28:     newindex++
29:     N = N - 1
30: end while
```

3.3. Measurement of Data Diversity

The equations for calculating F-Measure, G-Mean, and Q-Statistics can be seen in (1) until (9) [10] and [11]

$$\textit{True Negative Rate (TNrate)} = \frac{TN}{TN + FP} \quad (1)$$

$$\textit{False Negative Rate (F Nrate)} = \frac{FN}{TP+FN} \quad (2)$$

$$\textit{Positive Predictive Value (PPValue)} = \frac{TP}{TP + FP} \quad (3)$$

$$\textit{Negative Predictive Value (NPValue)} = \frac{TN}{TN + FN} \quad (4)$$

$$\underline{R}ecall = TPrate = \frac{TP}{TP + FN} \quad (5)$$

$$\underline{P}recision = PPValue = \frac{TP}{TP + FP} \quad (6)$$

$$F\text{-}Measure = \frac{2RP}{R+P} \quad (7)$$

$$G\text{-}Mean = \sqrt{TPrate\ .\ TNrate} \quad (8)$$

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \quad (9)$$

4. Experimental Process

4.1. Dataset Description

The data used in this research are Iris, Balanced Scale Weight & Distance Database, Haberman, Thyroid, New-Thyroid1, and New-Thyroid2 from KEEL-Dataset Repository and UCI Machine Learning Repository. The description about the Dataset can be seen in Table 1.

Table 1: Dataset Description

Dataset	#Ex	#Atts	(%Min;%Max)	IR
Iris	150	4	(33.33,66.67)	2
Haberman	306	3	(27.42,73.58)	2.68
New-thyroid2	215	5	(16.28,83.72)	5.14
New-thyroid1	215	5	(16.28,83.72)	5.14
New-thyroid	215	5	(13.95,86.05)	6.17
Balanced Scale	576	4	(7.84,92.16)	11.75

4.2. Testing

The experimental process is done using R Language. Testing of diversity data is done by measuring the value of F-Measure, G-Means, and Q-Statistics. The high F-Measure value indicates that the precision produced is good enough and the high G-Means value indicates that the balance of positive samples (minority class) and negative samples (majority class) is quite good. While the lower the value of Q-Statistics means the higher the value of diversity. The test results of F-Measure, G-Means, and Q-Statistics for Class Imbalance Handling in Clustering Result of Iris Dataset can be seen in Table 2.

Table 2: Testing Result of Iris Dataset

Testing Number	*SMOTEBoost*			HAR		
	F-Measure	*G-Means*	*Q-Statistics*	*F-Measure*	*G-Means*	*Q-Statistics*
1	0.72	0.83	0.42	0.929	0.93	0.15
2	0.82	0.79	0.37	0.89	0.89	0.33
3	0.79	0.81	0.44	0.86	0.86	0.06
4	0.81	0.77	0.39	0.88	0.88	0.1
5	0.8	0.79	0.45	0.88	0.88	0.25
6	0.69	0.75	0.47	0.76	0.76	0.28
7	0.79	0.74	0.43	0.67	0.75	0.11
8	0.82	0.83	0.39	0.86	0.86	0.12
9	0.82	0.8	0.4	0.88	0.87	0.27
10	0.8	0.76	0.36	0.81	0.78	0.047
Average	0.786	0.787	0.412	0.842	0.846	0.1717

The test results of F-Measure, G-Means, and Q-Statistics for Class Imbalance Handling in Clustering Result of Haberman Dataset can be seen in Table 3.

Table 3: Testing Result of Haberman Dataset

Testing Number	*SMOTEBoost*			HAR		
	F-Measure	*G-Means*	*Q-Statistics*	*F-Measure*	*G-Means*	*Q-Statistics*
1	0.43	0.56	0.46	0.42	0.54	0.12
2	0.43	0.49	0.52	0.55	0.64	0.41
3	0.52	0.58	0.48	0.56	0.65	0.28
4	0.44	0.52	0.45	0.53	0.61	0.31
5	0.43	0.56	0.39	0.53	0.62	0.31
6	0.42	0.54	0.42	0.62	0.7	0.4
7	0.55	0.52	0.53	0.59	0.67	0.55
8	0.44	0.49	0.52	0.63	0.7	0.5
9	0.34	0.45	0.34	0.53	0.63	0.24
10	0.38	0.43	0.4	0.56	0.65	0.5
Average	0.438	0.514	0.451	0.552	0.641	0.362

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 4.

Table 4: Testing Result of New-Thyroid2 Dataset

Testing Number	*SMOTEBoost*			HAR		
	F-Measure	*G-Means*	*Q-Statistics*	*F-Measure*	*G-Means*	*Q-Statistics*
1	0.41	0.49	0.59	0.74	0.87	0.33
2	0.5	0.48	0.42	0.8	0.81	0.94
3	0.42	0.46	0.52	0.65	0.77	0.36
4	0.4	0.48	0.57	0.7	0.83	0.76
5	0.47	0.49	0.58	0.67	0.79	0.44
6	0.4	0.52	0.57	0.8	0.81	0.57
7	0.43	0.5	0.53	0.86	0.88	0.23
8	0.42	0.53	0.56	0.65	0.8	0.15
9	0.48	0.51	0.5	0.79	0.81	0.94
10	0.47	0.52	0.51	0.74	0.84	0.57
Average	0.44	0.498	0.535	0.74	0.821	0.529

The test results of F-Measure, G-Means, and Q-Statistics for Class Imbalance Handling in Clustering Result of New-Thyroid1 Dataset can be seen in Table 5.

Table 5: Testing Result of New-Thyroid1 Dataset

Testing Number	*SMOTEBoost*			HAR		
	F-Measure	*G-Means*	*Q-Statistics*	*F-Measure*	*G-Means*	*Q-Statistics*
1	0.49	0.67	0.51	0.76	0.86	0.2
2	0.52	0.64	0.49	0.91	0.94	0.12
3	0.5	0.62	0.46	0.85	0.91	0.42
4	0.48	0.45	0.48	0.93	0.97	1
5	0.52	0.61	0.51	0.87	0.89	0.23
6	0.43	0.47	0.53	0.83	0.87	0.52
7	0.52	0.62	0.4	0.84	0.89	0.29
8	0.62	0.56	0.48	0.83	0.87	0.65
9	0.48	0.46	0.47	0.88	0.89	0.5
10	0.47	0.53	0.49	0.8	0.86	0.82
Average	0.503	0.563	0.482	0.85	0.895	0.475

The test results of F-Measure, G-Means, and Q-Statistics for Class Imbalance Handling in Clustering Result of New-Thyroid Dataset can be seen in Table 6.

Table 6: Testing Result of New-Thyroid Dataset

Testing Number	*SMOTEBoost*			HAR		
	F-Measure	*G-Means*	*Q-Statistics*	*F-Measure*	*G-Means*	*Q-Statistics*
1	0.49	0.55	0.49	0.67	0.79	0.265
2	0.51	0.52	0.52	0.73	0.76	0.295
3	0.47	0.54	0.42	0.48	0.64	0.086
4	0.61	0.63	0.37	0.75	0.79	0.66
5	0.53	0.56	0.48	0.44	0.67	0.044
6	0.52	0.54	0.51	0.62	0.73	0.43
7	0.49	0.55	0.56	0.79	0.8	0.85
8	0.39	0.5	0.39	0.58	0.71	0.205
9	0.49	0.47	0.46	0.43	0.57	0.43
10	0.5	0.48	0.45	0.53	0.64	0.82
Average	0.5	0.534	0.465	0.602	0.71	0.4085

The test results of F-Measure, G-Means, and Q-Statistics for Class Imbalance Handling in Clustering Result of Balance Scale Weight & Distance Database Dataset can be seen in Table 7.

Table 7: Testing Result of Balance Scale Weight & Distance Database Dataset

Testing Number	*SMOTEBoost*			HAR		
	F-Measure	*G-Means*	*Q-Statistics*	*F-Measure*	*G-Means*	*Q-Statistics*
1	0.48	0.5	0.67	0.65	0.65	0.25
2	0.51	0.52	0.57	0.63	0.66	0.008
3	0.5	0.53	0.67	0.57	0.58	0.86
4	0.49	0.51	0.48	0.7	0.75	0.59
5	0.51	0.48	0.65	0.53	0.53	0.89
6	0.46	0.54	0.47	0.48	0.49	0.47
7	0.4	0.46	0.5	0.62	0.56	0.77
8	0.53	0.48	0.56	0.46	0.51	0.97
9	0.52	0.54	0.58	0.82	0.81	0.38
10	0.5	0.48	0.52	0.66	0.67	0.39
Average	0.49	0.504	0.567	0.612	0.616	0.5578

5. Result and Discussion

The results showed that in general both the SMOTEBoost method and the HAR method can overcome the problem of class imbalance. When viewed from the side of data diversity which involves measurement of F-Measure, G-Means, and Q-Statistics then HAR Model also gives better result than SMOTEBoost method. This good data diversity is generated through the application of Different Contribution Sampling which effectively classifies both Majority and Minority Class into SV Sets and NSV Sets where processing both on NSV Sets from Majority Class and processing on SV Sets from Minority Class can provide Diversity data better.

6. Conclusion

The conclusion of this research are as follows. First, Hybrid Approach Redefinition (HAR) can handle class imbalance problem. Second, it is confirmed that Hybrid Approach Redefinition (HAR) can get better data diversity compared to SMOTEBoost
Our case study is using numerical datasets and in the future, it should be another study on non-numerical datasets and can apply it in handle class imbalance problem in Multi-Class Dataset. The importance of this research for future studies is that Diversity is essential in order to build an accurate ensemble of classifiers.

Acknowledgement

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

References

[1] Chawla NV, Japkowicz N & Kolcz A (2004), Special Issue Learning Imbalanced Datasets. SGIKDD Explor. Newsl 6(1), 1-6
[2] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H & Bing G (2017), Learning From Class-Imbalanced Data. Experts Systems with Application 73, 220-239
[3] Pastor J F D, Rodriguez J J, Osorio C I G & Kuncheva L I (2015), Diversity techniques improve the performance of the best imbalance learning ensembles. Information Sciences 325, 98-117
[4] Roy A, Cruz R M O, Sabourin M & Cavalcanti G D C (2018), A Study on combining Dynamic Selection and Data Preprocessing for Imbalance Learning. Neurocomputing
[5] Hartono, Sitompul O S, Tulus, Nababan E B (2018), Optimization Model of K-Means Clustering Using Artificial Neural Networks to Handle Class Imbalance Problem. IOP Conference Series: Materials Science and Engineering, 288, 012075.
[6] Galar M, Fernandez A, Barrenechea E & Bustince H (2012), 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(4), 1-21

[7] Jian C, Gao J & Ao Y (2016), A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble. Neurocomputing 193, 115-122
[8] Kuncheva L I, *Combining Pattern Classifiers*, John Wiley & Sons, (2004), pp. 295-327
[9] Wang S & Yao X, "Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models", *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining*, (2009)
[10] Sun Y, Kamel M S, Wong A K C & Wang Y (2007), Cost-Sensitive Boosting for Classification of Imbalanced Data. Pattern recognition 10, 3358-3378
[11] Yule G U (1900), On The Association of Attributes in Statistics. Philosophical Transactions of The Royal Society of London A194, 257-319
[12] Pastor J F D, Rodriguez J J, Osorio C I G, Kuncheva L I (2015), Random Balance: Ensembles of Variable Priors Classifiers for Imbalanced Data. Knowledge-Based Systems 85, 96-111
[13] Chawla N, Bowyer K, Hall L & Kegelmeyer P (2002), SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research 16, 321-357