PAPER NAME

4 Artikel IJAIN November 2021 Hartono.pdf

AUTHOR

Hartono

WORD COUNT

5622 Words

CHARACTER COUNT

31190 Characters

PAGE COUNT

12 Pages

FILE SIZE

651.9KB

SUBMISSION DATE

Feb 23, 2024 12:03 AM GMT+7

REPORT DATE

Feb 23, 2024 12:14 AM GMT+7

15% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 0% Publications database
- Crossref Posted Content database
- 15% Submitted Works database

Excluded from Similarity Report

- Internet database
- Crossref database
- Bibliographic material

Hybrid approach redefinition with cluster-based instance selection in handling class imbalance problem

Hartono [a,1,*], Erianto Ongko [b,2], Dahlan Abdullah [c,3]

[a] Department of Computer Science, Universitas Potensi Utama, Medan, Indonesia
[b] Department of Informatics, Akademi Teknologi Industri Immanuel, Medan, Indonesia
[c] Department of Informatics, Universitas Malikussaleh, Lhokseumawe, Indonesia
[1] hartonoibbi@gmail.com; [2] eriantoongko@gmail.com; [3] dahlan@unimal.ac.id
* corresponding author

ARTICLE INFO

Article history
Received June 24, 2020
Revised October 31, 2021
Accepted November 22, 2021
Available online November 30, 2021

Keywords
Class Imbalance
Hybrid approach redefinition
Hybrid ensembles
Classifier
Data diversity

ABSTRACT

Class Imbalance problems often occur in the classification process, the existence of these problems is characterized by the tendency of a class to have instances that are much larger than other classes. This problem certainly causes a tendency towards low accuracy in minority classes with smaller number of instances and also causes important information on minority classes not to be obtained. Various methods have been applied to overcome the problem of the imbalance class. One of them is the Hybrid Approach Redefinition method, one of the Hybrid Ensembles methods. The tendency to pay attention to the performance classifier has led to an understanding of the importance of selecting an instance that will be used as a classifier. In the classic Hybrid Approach Redefinition method classifier selection is done randomly using the Random Under Sampling approach, and it is interesting to study how performance is obtained if the sampling process is based on Cluster-Based by selecting existing instances. The purpose of this study is to apply the Hybrid Approach Redefinition method with Cluster-Based Instance Selection (CBIS) approach so that it can obtain a better performance classifier. The results showed that Hybrid Approach Redefinition with cluster-based instance selection gave better results on the number of classifiers, data diversity, and performance classifiers compared to classic Hybrid Approach Redefinition.

1. Introduction

Class Imbalance problems are characterized by the presence of a class with a number of instances that are much smaller (minority class) and other classes with a much larger number of instances (majority class) [1]. This problem is a major problem in the classification process and has attracted the attention of researchers in the fields of data mining and machine learning [2] and also often encountered in various classification problems [3]. If we discuss the problem of an imbalance class, the main consequence is low accuracy in minority classes, namely a class with a number of instances that are much smaller than majority classes, which is a class with a much larger number of instances [4] and the results obtained tend to be majority class [5]. Besides that, it should be noted that the classification process is carried out assuming that the distribution of instances in each class is the same, so if there is a problem of the imbalance class it will result in important information in a minority class with a much smaller number of instances that cannot be obtained [6] and it is also necessary to pay attention to the misclassification of minority classes [7].

Research on class imbalance problems has always been an interesting topic, especially if it is related to the problem of classification and machine learning which is very interesting to the attention of many researchers at this time. One method that draws the attention of researchers in overcoming the problem of this imbalance class is the hybrid approach [8]. Another approach that is widely used is data-driven and algorithm-driven [9]. The data-driven and algorithm-driven approach to handling imbalance classes experience the main issue of losing important information and training data overfitting. As for the hybrid ensembles, is training time [10]. To reduce training time, the hybrid ensembles method in principle adopts the sampling principle, which is combined with boosting [11][12]. However, other aspects need to be considered in handling the imbalance class, which is related to the number of classifiers and data diversity [13]. Another Hybrid Ensembles method, Hybrid Approach Redefinition, is one of the Hybrid Ensembles approaches based on sampling and boosting [14]. This method has been tested quite well in handling imbalance classes with a good number of classifiers and data diversity. However, the performance classifier needs to be considered, especially when faced with a dataset with a large number of attributes [15].

Sampling method is one of the approaches in handling class imbalance. This process is done by generating a new dataset from a dataset that has an imbalance class where the new dataset has a better distribution balance between majority and minority classes [16]. In general, the sampling method can be divided into three groups: Under Sampling, Over Sampling, and Hybrid Methods. The Under Sampling method focuses on reducing samples from Majority Class, while the Over Sampling method focuses on adding samples from Minority Class [10]. In Under Sampling, the instance selection process will get better results compared to ones trained using the original dataset [17]. In connection with the selection of samples in Under Sampling, it is known that cluster-based sample selection will get better results than random sample selection [18]. Another problem arises when handling two-class imbalance, with Under Sampling which has a primary focus on the Majority Class, causing the instance selection process to experience constraints because basically the selected instance is designed to distinguish groups of samples in multiclass datasets, it is difficult to apply to samples that only exist in one class, namely majority class [19]. Cluster-Based Instance Selection (CBIS) approach combines the Under Sampling method with the Instance Selection where the process instance selection will increase the ability of Under Sampling to Majority Class [20].

However, if handling imbalance classes only focus on the Majority Class, it will result in poor data diversity. At the same time, handling class imbalance is expected to use a small number of classifiers and obtain good data diversity [13]. Hybrid Approach Redefinition (HAR) Method offers handling class imbalance by using a small number of classifiers and good data diversity because handling focuses on the majority class and the minority class [14]. The Cluster-Based Instance Selection method is very interesting to be integrated with the Hybrid Approach Redefinition (HAR) Method. Especially in the process of Different Contribution Sampling using the Biased Support Vector Machine in the Majority Class [11]. Hybrid Approach Redefinition (HAR) Method with Cluster-Based Instance Selection is expected to obtain a smaller number of classifiers and better data diversity than the Classic Hybrid Approach Redefinition (HAR) Method.

Basically, the classification is based on the selection and placement of existing instances based on a number of existing classifiers [21]. This situation is the main thing that needs to be considered when dealing with the imbalanced class problem in a dataset with many attributes. Therefore, developing a Cluster-Based Instance Selection (CBIS) approach, which is an Under Sampling method, is stated to help well in the sampling process when there is a dataset with a large number of attributes [20]. The characteristics of clustering analysis with instance selection will complement each other in the Under Sampling process for majority classes. The performance classifier commonly used in research on the imbalance class is the measurement of Sensitivity, Specificity, F-Measure, and G-Mean [22]. Based on a number of previous studies, this study will discuss the application of the Hybrid Approach Redefinition method with the Cluster-Based Instance Selection (CBIS) approach in handling imbalance class problems so that a better performance classifier can be obtained, especially when compared to the Hybrid Approach Redefinition classic.

The rest of the paper is structured as follows. Section II presents the research method. Section III provides an experimental process using Hybrid Approach Redefinition with Cluster-Based Instance Selection and Hybrid Approach Redefinition Classic. The experimental process and dataset used are presented in Section IV with the results. Finally, Section V concludes the paper and gives recommendations for future works.

2. Method

The study was conducted to test the number of classifiers, data diversity, and performance classifier. Performance classifiers are measured based on sensitivity, specificity, F-Measure, and G-Means. In this study a comparison between Hybrid Approach Redefinition and Cluster-Based Instance Selection (CBIS) will be carried out with classic Hybrid Approach Redefinition. The process will begin with preprocessing stages, processing stages, and evaluation stages. The experimental process in this study will be carried out using datasets sourced from the KEEL Dataset Repository using datasets with many attributes [23].

2.1. Preprocessing Stage

This preprocessing stage will carry out the process of selecting instances that will be used as classifiers using the Cluster-Based Instance Selection (CBIS) Method and the SMOTEBoost Method. There is a slight difference with Hybrid Approach Redefinition classic, which uses Random Under Sampling and SMOTEBoost methods. The preprocessing stages in the Hybrid Approach Redefinition with CBIS can be seen in Fig. 1.

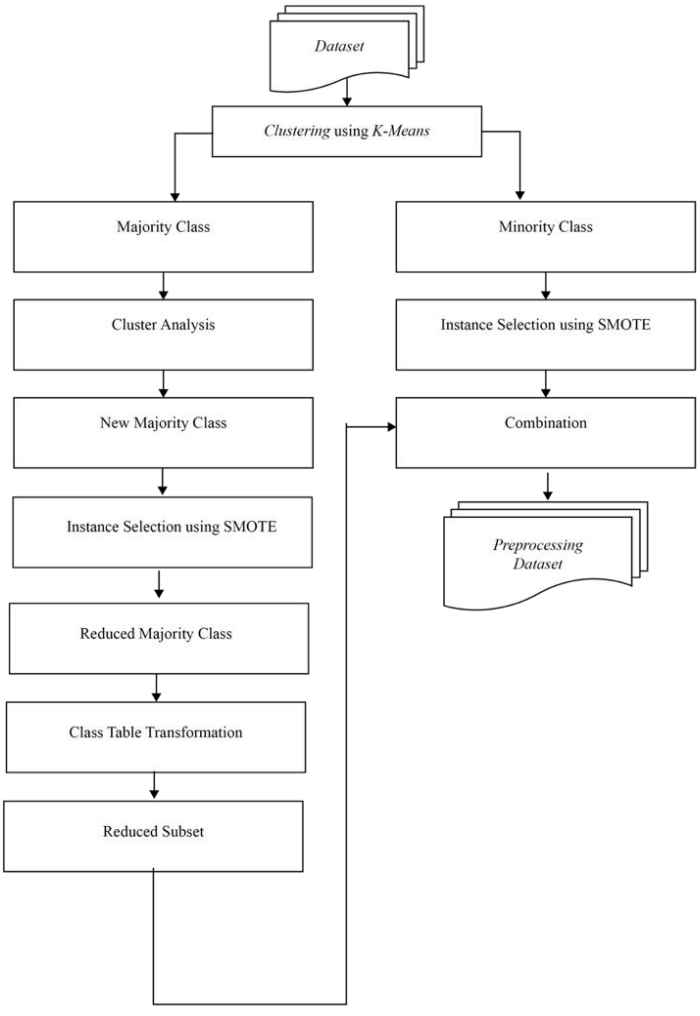

Fig. 1. Preprocessing Stage

In Fig. 1 can be seen that if the clustering results indicate the existence of an imbalance class problem, which is characterized by the existence of certain classes with a large number of instances (majority class) and the presence of classes with a small number of instances (minority class). A clustering analysis and instance selection process will be carried out in most classes. Clustering analysis will be conducted to group instances in majority classes where each instance belongs to a specific cluster, called a subclass of the majority class. Then, every instance that exists will associate with a new class label to generate a new majority class. The next stage will be the process of instance selection using SMOTEBoost, where this process is intended to measure the size of the classifier to produce a reduced majority class. Then the results will be transformed and recombined into a reduced subset of the majority class based on class label information. While the Minority class will undergo a cluster analysis stage for grouping instances that exist in minority classes and then will undergo the process of instance selection, the results will be combined with a reduced subset of the majority class to become a preprocessing dataset.

2.2. Processing Stage

The processing stage will be carried out using Biased Support Vector Machine. The Biased Support Vector Machine process will produce Support Vector Sets (SV Sets) and Non-Support Vector Sets (NSV Sets) for both majority class and minority class. For NSV Sets in Majority Class under sampling process will be carried out using Cluster-Based Instance Selection, while for SV Sets in Minority Class, instance selection will be processed using SMOTEBoost. The processing stages can be seen in Fig. 2.

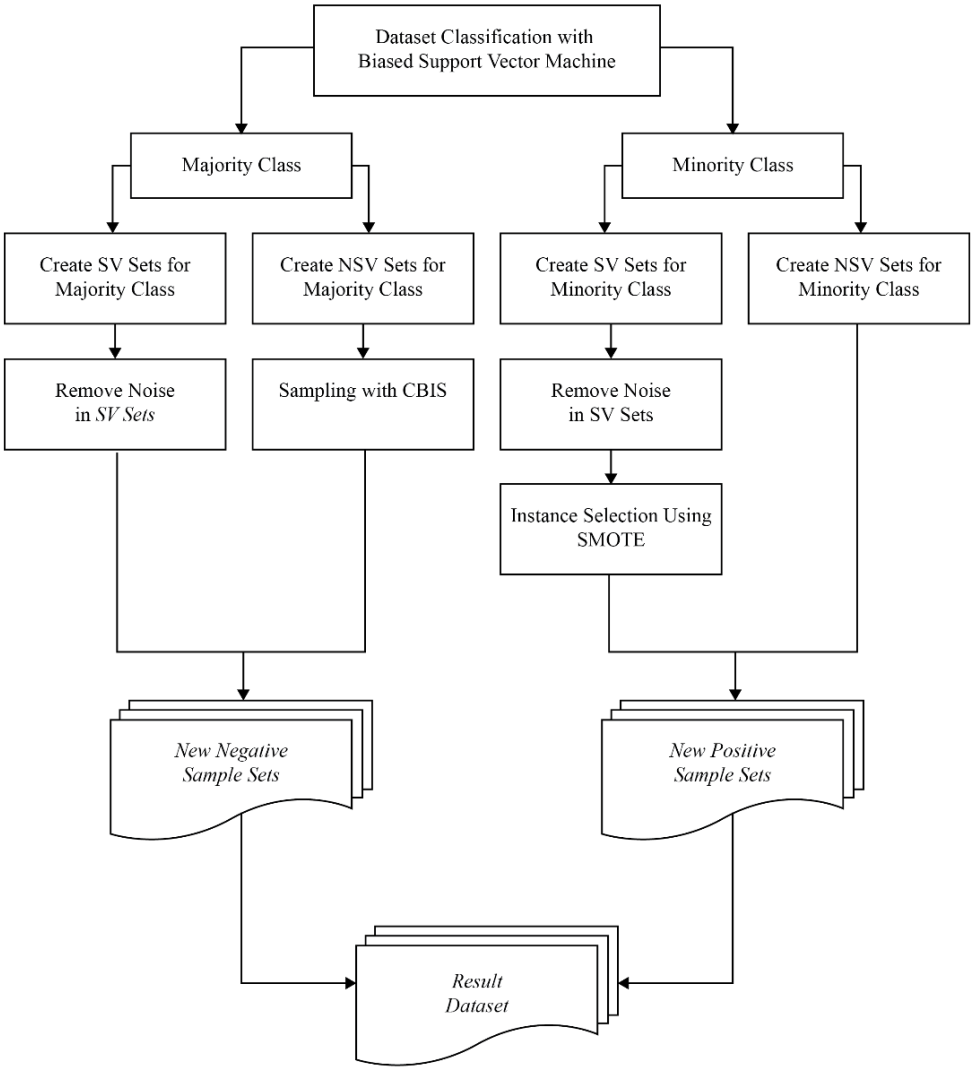

Fig. 2. Processing Stage

Preprocessing datasets originating from preprocessing stages will be classified using Biased Support Vector Machine to SV Sets and NSV Sets for Majority Class and Minority Class (Fig. 2). In the next process, the noise in the SV Sets Majority Class will be removed and then combined with the NSV Sets Majority Class which has undergone a sampling process using CBIS to become New Negative Sample Sets. In comparison, NSV Sets in the minority class will be combined with SV Sets in the majority class whose number has been removed and has undergone an instance selection process by using SMOTEBoost to become New Positive Sample Sets. The Result dataset will be both New Negative Sample Sets and New Positive Sample Sets.

2.3. Evaluation Stage

The evaluation stage is intended to compare the results obtained between Hybrid Approach Redefinition with Cluster-Based Instance Selection (CBIS) and classic Hybrid Approach Redefinition. Evaluation is done by looking at a number of parameters such as the number of classifiers, diversity data, and performance classifiers. The evaluation stages can be seen in Fig. 3.

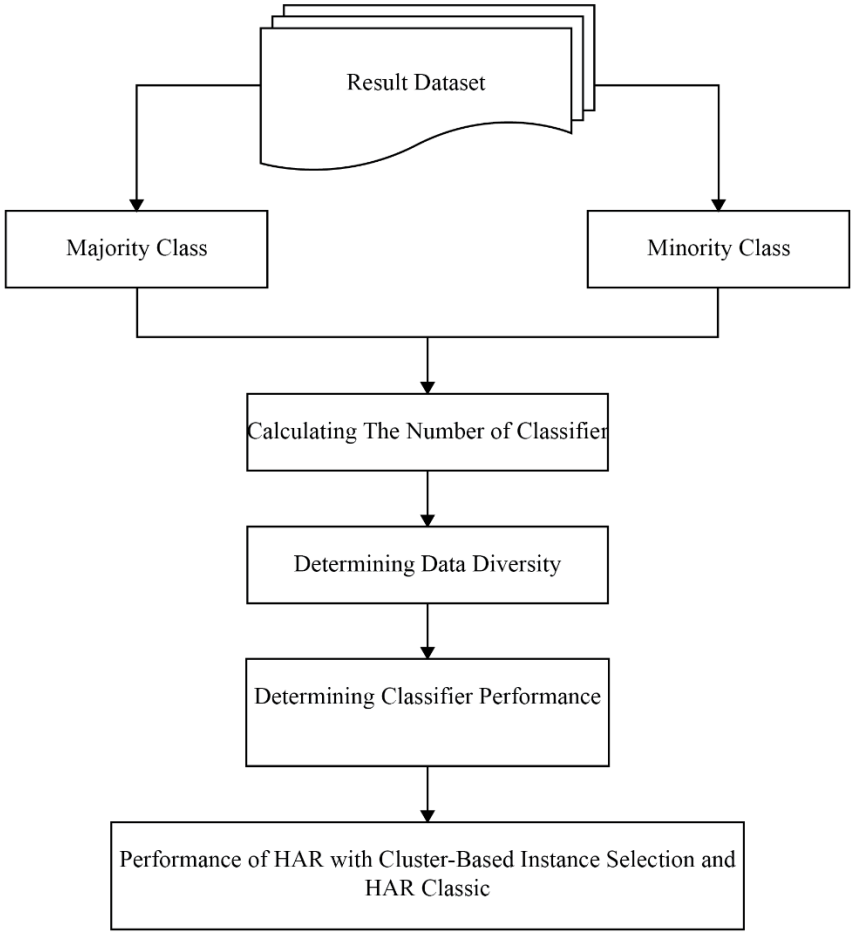

Fig. 3. Evaluation Stage

Based on Fig. 3, it can be seen that the processing dataset will measure the number of classifiers, data diversity, and performance classifier. The results are expected to illustrate the performance of handling the class imbalance between the results obtained by the Hybrid Approach Redefinition with Cluster-Based Instance Selection compared with classic Hybrid Approach Redefinition.

2.4. Under Sampling with Cluster-Based Instance Selection

The pseudocode of the Under Sampling process with Cluster-Based Instance Selection [20] shows as in Fig. 4. If there are problems with imbalance class, based on the CBIS method, cluster analysis and division will be carried out into several clusters in the majority class. Where will be given a specific class

label such as a class ID for each cluster that exists. Then the next step is to determine the instance of the majority class into the existing cluster. Then after that, an instance selection process will be carried out using the SMOTEBoost method. The process of this instance selection will produce an instance that has the best closeness with the majority class. After going through the process, a reduced subset that will undergo a noisy removal process produces S_Nonnoisy, an instance selected as majority class.

```
1:  Let S = Majority Class
2:  Delete ncol(S) as the class label of S
3:  Execute S to obtain the clusering list named AP
4:  Allocate each record of S to Size (AP)Clusters
5:  For (i = 1; i ≤ size (AP); i + +)
6:  {
7:     For (j = 1; j ≤ size (AP[i]); j + +)
8:     {
9:        Value = AP[i][j]
10:       S[Value, ncol(S)] = i
11:    }
12: }
13: Performance Instance Selection Over S to Produce Suvsets S_Noisy and S_NonNoisy
14: Replace ncol(S) of all records in S_nonnoisy with the Majority Class
15: Let RR = nrow (S_Noisy)/nrow(S)
16: Return S_Noisy and RR
```

Fig. 4. The pseudocode of the Under Sampling process with Cluster-Based Instance Selection

2.5. Hybrid Approach Redefinition with Cluster-Based Instance Selection

In the pseudocode of the Hybrid Approach Redefinition with Cluster-Based Instance Selection [20][24][25] can be seen if the classification results indicate a class imbalance problem (Fig. 5). The preprocessing stage begins with CBIS, which performs cluster analysis. At this stage, most classes will be divided into a number of specific clusters. Each existing instance is inserted into a particular cluster-specific so that each instance will be associated with a particular label class.

The formation of the New Majority Class will be based on the existing class label information. The next step is to process an instance selection using SMOTE, which will be done based on the generation of random numbers to move a number of instances from the majority class to the minority class. This is based on the level of closeness of the existing instance to the minority class. After this process is done, the results will be combined with the existing minority class to form a preprocessing dataset denoted by D '.

The next process will go into the processing stage, which will involve the Biased Support Vector Machine method, which will group both majority and minority classes into 2 (two) groups: SV Sets and NSV Sets. First, noise cleaning will be done on the SV Sets Majority Class and NSV Minority Class Sets. The next step is NSV Sets for majority classes to undergo a CBIS sampling process, which is then followed by the process of instance selection on the NSV Minority Class by using SMOTEBoost. This process will produce a New Majority Class and New Minority Class, which will then be combined to form the Result dataset.

```
Input: Total Size totalSize, Number of Majority S_N, Number of Minority S_P
totalSize ← |S|
S_N = {(x_i, y_i) ∈ S|y_i = −1}
S_P = {(x_i, y_i) ∈ S|y_i = +1}
majoritySize ← |S_N|
minoritySize ← |S_P|
Preprocessing Stage:
Execute S_N to obtain the clusering list named AP
Allocate each record of S_N to Size (AP)Clusters
For (i = 1; i ≤ size (AP); i + +)
{
  For (j = 1; j ≤ size (AP[i]); j + +)
  {
    Value = AP[i][j]
    S[Value, ncol(S)] = i
  }
}
k = Number of Nearest Neighbors
numattrs = number of attributes
Sample[ ][ ]: Minority Class Sample
DMajority_Reduced = Array of Majority
DMinority = Array of Minority
For (i = 1; i ≤ majoritySize; i + +)
{
  Compute k nearest neighbors
  Populate (N, i, nnarray)
}
While N ≠ 0 do
{
  for (i = 1; i ≤ numattrs; i + +)
  {
    dif[i] = sample[nnarray[i][attr] − sample[i][attr]
  }
}
For (i = 1; i ≤ majoritySize; i + +)
{
  Sort sample[i][attr]according to dif[i]
}
majoritySize_reduced = random number 1 to majoritySize
For (i = 1; i ≤ majoritySize ; i + +)
{
  if i ≤ majoritySize_reduced
    DMajority_Reduced[i][attr] = sample[i][attr]
  else
    DMinority[i][attr] = sample[i][attr]
Combine DMajority_Reduced with D_Minority become D'
Processing Stage:
T = number of Iteration
for(i = 1; i ≤ T; i + +)
{
  Classifying D'using B − SVM
  Identifying Majority Class
  Identifying Minority Class
  While (!endOfMajorityClass) do
  {
    NewSVSets ← Deleting Noise of SVSets
    NewNSVSets ← Sampling NSVSets using CBIS
  }
  While (!endOfMinorityClass) do
  {
    NewSVSets ← Deleting Noise of SVSets and Instance Selection using SMOTE
  }
}
Result Dataset
```

Fig. 5. Pseudocode of the Hybrid Approach Redefinition with Cluster-Based Instance Selection

2.6. Data Diversity

Data diversity is intended to measure the performance of a classifier, especially in situations where misclassification occurs. Misclassification is an unavoidable thing in handling imbalance classes. However, on the other hand, a small amount of misclassification is not necessarily good because when faced with a situation where the number of instances in a minority class is very small, it is easy to group all instances into majority classes. Therefore, good diversity data shows that misclassification does not only occur in one class; if there is misclassification, it should also be covered by merging with another classifier [26][27].

Suppose that Z= $\{z_1, \ldots, z_n\}$ which is a dataset that is in the decision region $\Re^n$, so that $z_j \in \Re^n$ it is an instance involved in the classification problem. Then the output of the classifier D_i as a classifier paired comparison matrix (relationship pairwise classifier) can be seen in Table 1.

Table 1. Relationship Pairwise Classifier Matrix

	D_k Correct (1)	D_k Wrong (0)
D_i Correct (1)	N^{11}	N^{10}
D_i Wrong (0)	N^{01}	N^{10}

Diversity data can be calculated using Q-Statistics [28] as in (1).

$$Qi,k = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \quad (1)$$

2.7. Classifier Performance

On Binary Class issues, positive samples refer to minority class, and negative samples refer to majority class. For the general classification results, the classification results can be grouped into 4 (four) groups, namely: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), and can be presented in the Confusion Matrix as can be seen in Table 2 [29].

Table 2. Confusion Matrix for A Binary Class Problem

		Predicted (Classified) as	
		Positive	***Class Negative***
Actually (Really is)	***Positive Samples***	True Positive (TP)	False Negative (FN)
	Negative Samples	False Positive (FP)	True Negative (TN)

For the measurement of performance classifier based on the confusion matrix, which can be seen in Table 2, the measurement is done based on several things as follows: 1) True Positive (TP) states the number of positive samples classified correctly as positive; 2) True Negative (TN) states the number of negative samples classified correctly as negative; 3) False Positive (FP) states the number of negative samples is classified incorrectly as positive; 4) False Negative (FN) states the number of positive samples classified incorrectly as negative.

The classifier performance that can be measured based on the confusion matrix [29] is:

1) Sensitivity states the ability of the classifier to identify positive samples correctly. The best value for sensitivity is 1 and the lowest value is 0. The sensitivity can be measured using (2).

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

2) Specificity states the ability of the classifier to identify the negative sample correctly. The best value for specificity is 1 and the lowest value is 0. The specificity can be measured using (3).

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

3) F-Measure usually refers to the harmonious average value between Precision and Recall. Precision states how well the classifier avoids the misclassification of the negative class as a positive class, and recall states how well the classifier classifies the positive class

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F - Measure = \frac{2RP}{R+P} \quad (6)$$

4) G-Mean states the classifier's ability to balance the accuracy of the classification on positive and negative samples.

$$G - Mean = \sqrt{Sensitivity \,.\, Specificity} \quad (7)$$

3. Result and Discussion

3.1. Dataset Description

The dataset used in this research are Page-Blocks, Vowel, and Vehicle1. The description of the dataset can be seen in Table 3. The dataset selected is a dataset with a large number of attributes. In general, the datasets represent the number of instances that vary: small, medium and large and also with small, medium, and large imbalance ratios.

Table 3. Dataset Description

Dataset	#Ex	#Atts	(%Min;%Max)	IR
Page-Blocks	5472	10	(10.23,89.77)	8.77
Vowel	988	13	(9.01,90.99)	10.10
Vehicle1	846	18	(28.37,71.63)	2.52

3.2. Testing

Tests will be conducted to obtain a performance picture of the Hybrid Approach Redefinition method with Cluster-Based Instance Selection and Classic Hybrid Approach Redefinition, especially in terms of the number of classifiers, diversity data, and also the performance classifier. The test will be carried out ten times for each method. The test results for the number of classifier and diversity data can be seen in Table 4.

Table 4. Testing Result for Number of Classifier and Data Diversity of Each Method

Dataset	Hybrid Approach Redefinition		Hybrid Approach Redefinition with Cluster-Based Instance Selection	
	Number of Classifier	*Data Diversity*	*Number of Classifier*	*Data Diversity*
Page-Blocks	517.1	0.916	512	0.878
Vowel	214.1	0.264	207.2	0.259
Vehicle1	207.1	0.615	206	0.595

For classifiers and data diversity, Hybrid Approach Redefinition with Cluster-Based Instance Selection can provide better results than the classic Hybrid Approach Redefinition; however, if it is seen that diversity data for Page-Blocks in both methods is still not good. Classic Hybrid Approach Redefinition gives results of 0.916, and Hybrid Approach Redefinition with Cluster-Based Instance

Selection gives a result of 0.878. Even though Hybrid Approach Redefinition with Cluster-Based Instance Selection still provides better results than Hybrid Approach Redefinition, the results obtained should be better. However, the data diversity results are quite good for a few instances.

The measurement results for the performance classifier based on sensitivity, specificity, F-Measure, and G-Mean can be seen in Table 5. In general, the performance classifier measurements show that Hybrid Approach Redefinition with Cluster-Based Instance Selection gives better results compared to classic Hybrid Approach Redefinition. Measurements for sensitivity, specificity, and F-Measure tend to show Hybrid Approach Redefinition with Cluster-Based Instance Selection to give better results. Whereas for the G-Mean, the results obtained are not much different, and for the Dataset and Vehicle1, the results given by the two methods are the same. The measurement for G-Mean given by both methods is good, and this means that the balance of predictive accuracy for the majority and minority classes is quite good.

Table 5. Testing Result for Sensitivity, Specificity, F-Measure, and G-Mean of Each Method

Dataset	Hybrid Approach Redefinition				Hybrid Approach Redefinition with Cluster-Based Instance Selection			
	Sensitivity	*Specificity*	*F-Measure*	*G-Mean*	*Sensitivity*	*Specificity*	*F-Measure*	*G-Mean*
Page-Blocks	0.521	0.918	0.542	0.691	0.542	0.921	0.567	0.71
Vowel	0.537	0.763	0.632	0.64	0.529	0.771	0.651	0.64
Vehicle1	0.467	0.496	0.64	0.481	0.471	0.492	0.701	0.481

Based on a series of tests in terms of the number of classifiers, Hybrid Approach Redefinition with Cluster-Based Instance Selection can reduce the number of classifiers, but not too much different. This is because classic Hybrid Approach Redefinition, in general, has been able to overcome the problem of the imbalance class with a very good number of classifiers. The sampling process in classic Hybrid Approach Redefinition that uses Random Under Sampling gives only slightly worse results than under sampling using Cluster-Based Instance Selection. However, what needs to be paid attention to is the data diversity. There is a tendency that the two methods have not been very effective in the dataset with a large number of instances. This means that there is a tendency for misclassification to occur in only one classifier group. This might be overcome by selecting a more appropriate instance selection method. Soleymani *et al.* [30] stated that SMOTEBoost tends to have weaknesses in providing good data diversity. The measurement results for sensitivity, specificity, F-Measure, and G-Mean given are very good. So there is no concern that there is a high number of misclassification in the minority class and majority class. A good accuracy balance is shown in the G-Mean value, which tends to be the same in the Hybrid Approach Redefinition with Cluster-Based Instance Selection and Classic Hybrid Approach Redefinition.

This study shows that the Hybrid Approach Redefinition with Cluster-Based Instance Selection provides better results than the classic Hybrid Approach Redefinition, both for the number of classifiers, diversity data, and performance classifiers. Future research is expected to focus on the problem of data diversity, especially for datasets with a large number of instances.

4. Conclusion

The study implemented Hybrid Approach Redefinition with Cluster-Based Instance Selection in handling class imbalance problem. The results showed that Hybrid Approach Redefinition with cluster-based instance selection gave better results on the number of classifiers, data diversity, and performance classifiers compared to classic Hybrid Approach Redefinition. This research discusses handling class imbalance for two-class imbalance problems, and future research can develop this method to handle Multiclass imbalance problems.

Acknowledgment

The authors thank the Directorate of Research and Development, under the Ministry of Education, Culture, Research, and Technology, Indonesia, for supporting this research.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. This work was supported by the Directorate of Research and Development, under the Ministry of Education, Culture, Research, and Technology, Indonesia.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] A. de Haro-García, G. Cerruela-García, and N. García-Pedrajas, "Ensembles of feature selectors for dealing with class-imbalanced datasets: A proposal and comparative study," *Inf. Sci. (Ny).*, 2020, doi: 10.1016/j.ins.2020.05.077.

[2] C. K. Maurya, D. Toshniwal, and G. V. Venkoparao, "Online anomaly detection via class-imbalance learning," in *2015 8th International Conference on Contemporary Computing, IC3 2015*, 2015, doi: 10.1109/IC3.2015.7346648.

[3] B. Richhariya and M. Tanveer, "A reduced universum twin support vector machine for class imbalance learning," *Pattern Recognit.*, 2020, doi: 10.1016/j.patcog.2019.107150.

[4] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," 2009, doi: 10.1007/978-0-387-09823-4_45.

[5] D. Tomar and S. Agarwal, "Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes," *Adv. Artif. Neural Syst.*, 2015, doi: 10.1155/2015/265637.

[6] Z. P. Agusta and Adiwijaya, "Modified balanced random forest for improving imbalanced data prediction," *Int. J. Adv. Intell. Informatics*, 2019, doi: 10.26555/ijain.v5i1.255.

[7] B. Richhariya and M. Tanveer, "EEG signal classification using universum support vector machine," *Expert Syst. Appl.*, 2018, doi: 10.1016/j.eswa.2018.03.053.

[8] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," 2016, doi: 10.1007/s13748-016-0094-0.

[9] C. K. Maurya and D. Toshniwal, "Large-Scale Distributed Sparse Class-Imbalance Learning," *Inf. Sci. (Ny).*, 2018, doi: 10.1016/j.ins.2018.05.004.

[10] 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," 2012, doi: 10.1109/TSMCC.2011.2161285.

[11] C. Jian, J. Gao, and Y. Ao, "A new sampling method for classifying imbalanced data based on support vector machine ensemble," *Neurocomputing*, 2016, doi: 10.1016/j.neucom.2016.02.006.

[12] 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.*, 2015, doi: 10.1016/j.eswa.2014.08.025.

[13] P. Yang, P. D. Yoo, J. Fernando, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, "Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications," *IEEE Trans. Cybern.*, 2014, doi: 10.1109/TCYB.2013.2257480.

[14] Hartono, E. Ongko, O. S. Sitompul, Tulus, E. B. Nababan, and D. Abdullah, "Hybrid Approach Redefinition (HAR) Method with Loss Factors in Handling Class Imbalance Problem," in *Proceeding - 2018 International Symposium on Advanced Intelligent Informatics: Revolutionize Intelligent Informatics Spectrum for Humanity, SAIN 2018*, 2019, doi: 10.1109/SAIN.2018.8673370.

[15] Hartono, O. S. Sitompul, Tulus, and E. B. Nababan, "Biased support vector machine and weighted-SMOTE in handling class imbalance problem," *Int. J. Adv. Intell. Informatics*, 2018, doi: 10.26555/ijain.v4i1.146.

[16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, 2009, doi: 10.1109/TKDE.2008.239.

[17] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Mach. Learn.*, 2000, doi: 10.1023/A:1007626913721.

[18] V. Vigneron and H. Chen, "A multi-scale seriation algorithm for clustering sparse imbalanced data: application to spike sorting," *Pattern Anal. Appl.*, 2016, doi: 10.1007/s10044-015-0458-2.

[19] S. García, J. Derrac, J. R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, doi: 10.1109/TPAMI.2011.142.

[20] C. F. Tsai, W. C. Lin, Y. H. Hu, and G. T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Inf. Sci. (Ny).*, 2019, doi: 10.1016/j.ins.2018.10.029.

[21] I. C. Irsan and M. L. Khodra, "Hierarchical multi-label news article classification with distributed semantic model based features," *Int. J. Adv. Intell. Informatics*, 2019, doi: 10.26555/ijain.v5i1.168.

[22] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, 2019, doi: 10.1016/j.patcog.2019.02.023.

[23] J. Alcalá-Fdez *et al.*, "KEEL: A software tool to assess evolutionary algorithms for data mining problems," *Soft Comput.*, 2009, doi: 10.1007/s00500-008-0323-y.

[24] Hartono, O. S. Sitompul, E. B. Nababan, Tulus, D. Abdullah, and A. S. Ahmar, "A new diversity technique for imbalance learning ensembles," *Int. J. Eng. Technol.*, 2018, doi: 10.14419/ijet.v7i2.11251.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, 2002, doi: 10.1613/jair.953.

[26] J. F. Díez-Pastor, J. J. Rodríguez, C. I. García-Osorio, and L. I. Kuncheva, "Diversity techniques improve the performance of the best imbalance learning ensembles," *Inf. Sci. (Ny).*, 2015, doi: 10.1016/j.ins.2015.07.025.

[27] L. I. Kuncheva, *Combining Pattern Classifiers*, 2004, doi: 10.1002/0471660264.

[28] G. U. Yule, "VII. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c," *Philos. Trans. R. Soc. London. Ser. A, Contain. Pap. a Math. or Phys. Character*, vol. 194, no. 252–261, pp. 257–319, Jan. 1900, doi: 10.1098/rsta.1900.0019.

[29] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *Int. J. Adv. Soft Comput. its Appl.*, 2015.

[30] 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, doi: 10.1016/j.eswa.2018.01.023.

15% Overall Similarity

Top sources found in the following databases:

- 0% Publications database
- Crossref Posted Content database
- 15% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	School of Business and Management ITB on 2023-03-05 Submitted works	5%
2	University of Birmingham on 2020-09-17 Submitted works	1%
3	Birkbeck College on 2018-09-11 Submitted works	<1%
4	SDM Universitas Gadjah Mada on 2023-03-29 Submitted works	<1%
5	University of Portsmouth on 2015-04-23 Submitted works	<1%
6	University of Computer Studies on 2020-09-17 Submitted works	<1%
7	SUNY, Binghamton on 2012-12-10 Submitted works	<1%
8	University of Computer Studies on 2019-03-26 Submitted works	<1%
9	University of Glamorgan on 2023-02-01 Submitted works	<1%

10	The Robert Gordon University on 2021-02-24 Submitted works	<1%
11	iGroup on 2015-08-24 Submitted works	<1%
12	Loughborough University on 2021-09-01 Submitted works	<1%
13	Loughborough University on 2022-09-05 Submitted works	<1%
14	Cranfield University on 2019-08-21 Submitted works	<1%
15	SDM Universitas Gadjah Mada on 2023-04-12 Submitted works	<1%
16	Unviersidad de Granada on 2018-02-09 Submitted works	<1%
17	National University of Singapore on 2011-09-22 Submitted works	<1%
18	Universiti Sultan Zainal Abidin on 2015-02-18 Submitted works	<1%
19	Auckland University of Technology on 2011-11-01 Submitted works	<1%
20	University of Newcastle upon Tyne on 2020-02-01 Submitted works	<1%
21	Erasmus University of Rotterdam on 2022-03-22 Submitted works	<1%

22	School of Business and Management ITB on 2018-09-06 Submitted works	<1%
23	Eastern Mediterranean University on 2023-07-21 Submitted works	<1%
24	Multimedia University on 2017-12-29 Submitted works	<1%
25	Politeknik Statistika STIS on 2023-09-07 Submitted works	<1%
26	Swinburne University of Technology on 2020-06-12 Submitted works	<1%
27	The University of Manchester on 2021-09-16 Submitted works	<1%
28	Universitas Mulawarman on 2020-03-16 Submitted works	<1%
29	Universitas Pendidikan Indonesia on 2024-02-22 Submitted works	<1%
30	University of Strathclyde on 2016-09-01 Submitted works	<1%
31	University of Ulster on 2021-01-21 Submitted works	<1%
32	University of Ulster on 2023-10-02 Submitted works	<1%
33	University of Leeds on 2019-09-05 Submitted works	<1%

34	University of Sheffield on 2015-05-12 Submitted works	<1%
35	Universidad de Salamanca on 2019-06-13 Submitted works	<1%