Combining feature selection and hybrid approach redefinition in handling class imbalance and overlapping for multi-class imbalanced

Hartono[1], Erianto Ongko[2], Yeni Risyani[3]
[1]Department of Computer Science, STMIK IBBI, Medan, Indonesia
[1]Department of Computer Science, Universitas Potensi Utama, Medan, Indonesia
[2]Department of Informatics, Akademi Teknologi Industri Immanuel, Medan, Indonesia
[3]Department of Information Systems, STMIK IBBI, Medan, Indonesia

Article Info

Article history:

Received Oct 7, 2020
Revised Dec 10, 2020
Accepted Dec 26, 2020

Keywords:

Class imbalance
Feature selection
Hybrid approach redefinition
Multi-class imbalanced
Overlapping

ABSTRACT

In the classification process that contains class imbalance problems. In addition to the uneven distribution of instances which causes poor performance, overlapping problems also cause performance degradation. This paper proposes a method that combining feature selection and hybrid approach redefinition (HAR) Method in handling class imbalance and overlapping for multi-class imbalanced. HAR was a hybrid ensembles method in handling class imbalance problem. The main contribution of this work is to produce a new method that can overcome the problem of class imbalance and overlapping in the multi-class imbalance problem. This method must be able to give better results in terms of classifier performance and overlap degrees in multi-class problems. This is achieved by improving an ensemble learning algorithm and a preprocessing technique in HAR using minimizing overlapping selection under SMOTE (MOSS). MOSS was known as a very popular feature selection method in handling overlapping. To validate the accuracy of the proposed method, this research uses augmented R-Value, Mean AUC, Mean F-Measure, Mean G-Mean, and Mean Precision. The performance of the model is evaluated against the hybrid method (MBP+CGE) as a popular method in handling class imbalance and overlapping for multi-class imbalanced. It is found that the proposed method is superior when subjected to classifier performance as indicate with better Mean AUC, F-Measure, G-Mean, and precision.

Corresponding Author:

Erianto Ongko
Department of Informatics
Akademi Teknologi Industri Immanuel
Jalan Jenderal Gatot Subroto Nomor 325, Medan, Indonesia
Email: eriantoongko@gmail.com

1. INTRODUCTION

The problem of class imbalance is one of the interesting problems and is included in the top 10 challenging problems, especially when we discuss classification. Where, this problem comes from the existence of one class with a much higher number of instances (majority class) compared to other classes (minority classes) [1]. Majority classes are often referred to as negative classes and minority classes are positive classes. This term arises because the minority class sometimes contains information that is important to observe even though it is often overlooked because the classification results tend to give poor accuracy to classes with smaller number of instances [2]. However, in addition to the class imbalance there are other problems that need attention due to the accuracy of the classification process. These problems often go unnoticed. The problem is overlapping [3]. Surely the problem of overlapping is not a new problem. An instance of a class is said to be in the overlapping area if the value of k nearest neighbors (KNN), is too close to another class (greater than the value of h which is often assumed to be k / 2 [4]. This problem can decrease accuracy more if compared to the imbalance class.

Research on overlapping has not attracted much attention, compared to the class imbalance problem that has caught the attention of many [5]. When these two problems combine, the problem that will occur becomes more serious and will become more difficult if the problem occurs in a multi-class dataset [6]. Class imbalance and overlapping problems are relatively easier to handle in binary class problems and will be more difficult to handle in multi-class problems [7] Some researchers consider that feature selection is the best method in dealing with overlapping.

Research conducted by a number of researchers indeed shows that feature selection offers many advantages in overcoming class imbalance problems, especially those involving the elimination of uninformative predictors, reducing the dimensions of feature space, and most importantly feature selection can also overcome class imbalance problems [8, 9]. There are a number of feature selection methods that can be used in overcoming overlapping problems, including selection under no sampling [10, 11] and selection under SMOTE [12]. Has propsosed the minimizing overlapping selection under no-sampling (MOSNS) and minimizing overlapping selection under SMOTE (MOSS) methods in overcoming the overlapping problem and both methods have relatively the same performance [5, 13]. However, the class imbalance problem cannot be overcome by simply using the feature selection. The process of prepocessing in the form of data training resampling is absolutely necessary [14]. The best method of handling imbalance classes in conditions that allow for a preprocessing process to overcome the weaknesses of feature selection is the hybrid ensembles method [15].

In fact, a number of studies on handling imbalance classes and overlapping binary classes and multi-classes have been conducted by a number of researchers. Has proposed a hybrid method using a modified back-propagation and gabriel graph editing (GCE) on the class imbalance and overlapping problems for multi-class problems and get the results that the methods they propose, get good results, but if applied to highly imbalanced datasets, the results obtained are still not good and if coupled with the application of SMOTE, the results obtained can be better [16]. Research conducted by also provides results that SMOTE still provides the best solution [6]. Has defined the importance of the sampling process in the preprocessing stage by using clustering based under sampling (CLUSBUS) [17]. The Oversampling method by using SMOTE has also been used by [18] in dealing with class imbalance and overlapping problems.

It can be seen that most of the research that deals with class imbalance and overlapping problems is in the binary class problem and there are not many studies that discuss multi-class problems. One of the studies directly related to multi-class problems is the research conducted by [16] the method that they have proposed has been able to overcome the class imbalance and overlapping problems. However, the method they propose is still experiencing limitations when viewed from the performance classifier especially those relating to AUC and G-Mean. Research conducted by [16], has contributed thought in the form of the importance of hybrid methods that use the feature selection process at the preprocessing stage.

This paper proposes a method that combines feature selection and hybrid approach redefinition (HAR) Method in handling class imbalance and overlapping for multi-class imbalanced. HAR is a Hybrid Ensembles method in dealing with class imbalanced problems, where this method uses SMOTE [19] as a preprocessing stage [20]. The Feature selection method used is MOSS [5] and will be used as a preprocessing stage in this study replacing the SMOTE method that was previously used in the HAR Method, specifically this is intended to develop the HAR Method capability in overlapping handling. The combination of Feature Selection by using MOSS with HAR Method is intended to obtain good results in handling class imbalance and overlapping in multi-class imbalanced. The results obtained will be compared with the MBP + GCE method which is one of the excellent methods in handling class imbalance and overlapping in multi-class imbalanced. Comparison of these results was observed using augmented R-value, mean AUC, mean F-measure, mean G-Mean, and mean precision.

2. RELATED WORKS

2.1. Augmented R-Value for Multi-Class

The R-value of each class shows the portion of an instance that overlaps the area. Research conducted by [20] shows that R-value has a close relationship with performance classifier. Has proposed a method for determining R-Value for multi-class problems as can be seen in (1) [16]:

$$R_{aug}(D[V]) = \frac{\sum_{i=0}^{k-1}|C_{k-1-i}|R(C_i)}{\sum_{i=0}^{k-1}|C_i|} \tag{1}$$

Where $C_0, C_1, \ldots, C_{k-1}$ are k class labels with $|C_0| \geq |C_1| \geq \cdots \geq |C_{k-1}|$, $D[V]$: Dataset D containing predictors in set V, and $R(C_i)$ is the ratio of instance C_i that are located in overlapping area with all different categories C_j. Larger R_{Aug} is higher overlap degree of a dataset.

2.2. Hybrid ensembles

The algorithm for creating a hybrid ensemble is as follows [21]:

```
$Input: A\ dataset\ D, a\ se\ of\ classification\ algorithms\ G, the\ number\ of\ classifiers\ n$
$Output: An\ ensemble\ C$
    $Steps:$
$(1) For\ i = 1\ to\ n$
$(2)\ Use\ bootstrap\ sampling\ to\ sample\ D\ and\ Generate\ T_i, whics\ is\ of\ the\ same\ size\ of\ D$
$(3)\ Select\ the\ ([i\ modulo\ |G|] + 1)th\ element\ in\ G\ as\ A_i$
$(4)\ Train\ C_i\ by\ applying\ A_i\ on\ T_i$
$(5) End\ For$
$(6) Return\ C = \cup_{i=1}^{n} C_i$
```

In the previous algorithm it can be seen that in hybrid ensembles the process starts from determining bootstrap sampling at the preprocessing stage. In general, the sampling method used is SMOTE. Then it will proceed with the processing stage by selecting the appropriate classification algorithm. One of the Hybrid Ensembles methods is hybrid approach redefinition (HAR) which gives excellent results in handling class imbalance. In the original form of HAR the preprocessing stage is carried out using the random balance ensemble method and the processing stage is carried out using different contribution sampling [22]. To deal with class imbalance problems as well as overlapping problems in multi-class problems, the preprocessing technique will be modified using Filter Selection, namely minimizing overlapping selection under SMOTE (MOSS).

2.3. Minimizing overlapping selection under SMOTE (MOSS)

The algorithm for MOSS is as follows [5]:

```
$1: X - matrix\ with\ p\ predictors: X = [x_1, x_2, \ldots, x_p]; class\ label: y$
$2: Over - sampling\ the\ Minority\ Class\ with\ SMOTE; merging\ the\ generated\ data\ with\ original\ ones\ to\ get\ updated\ X - martix, X_{new}\ and\ updated\ class\ label\ Y_{new}$
$3: X \leftarrow X_{new}; Y \leftarrow Y_{new}$
$4: Establish\ sparse\ regularization\ path\ \hat{\beta}(\lambda, \alpha) according\ to\ equation\ 2$
$5: Compute\ the\ optimal\ (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T\ via\ the\ equation\ 3$
$6: Select\ those\ feature\ with\ \hat{\beta}_j \neq 0\ for\ j = 1, 2, \ldots, p$
```

Sparse selection that would be used to establish sparse regulatization can be seen in (2) [23]:

$$C_a(\beta) = \frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \tag{2}$$

where $||\beta||\ is\ lasso\ penalty$ and $C_a(\beta)$ is lasso penalty of each instance and $\alpha\ \epsilon\ [0,1]$.

Loss Penalties that would be used to compute the optimal $\hat{\beta}_j$ can be seein in (3):

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}(y_i\beta^T x_i - \ln(1 + \beta^T x_i)) \tag{3}$$

Where the $Loss$ are the Loss from sparse logistic regression.

In pseudocode, it can be seen that MOSS begins with determining a number of predictor matrices and determining the class label used. After that, the process will continue with the process of over sampling the minority classes using SMOTE and will be continued by updating the predictor matrix which contains a number of classifiers and also updating a new class label for each instance. The process will be continued with the process of determining sparse regularization based on the lasso penalty value from each instance and then the loss penalties will be determined as the basis for determining the optimal conditions. The loss penalties originate from sparse logistic regression. The process will continue with determining the optimal conditions based on the resulting loss value. The preprocessing stage in Hybrid Ensembles will be carried out using the MOSS method so that the preprocessed dataset is generated.

2.4. Classifier performance

Confusion Matrix shows the outcome of the classification results for each instance as can be seen in Table 1. [24].

Table 1. Confusion matrix

	Classified as Positive	Classified as Negative
Positive Samples	True Positive (TP)	False Negative (FN)
Negative Samples	False Positive (FP)	True Negative (TN)

The performance classifier measurement uses a number of parameters as follows:

$$TPR=\frac{\mathrm{TP}}{\mathrm{TP + FN}} \tag{4}$$

$$FPR=\frac{\mathrm{FP}}{\mathrm{TN+FP}} \tag{5}$$

$$TNR=\frac{\mathrm{TN}}{\mathrm{TN + FP}} \tag{6}$$

$$Recall= TPR \tag{7}$$

$$Precision= PPValue =\frac{\mathrm{TP}}{\mathrm{TP + FP}} \tag{8}$$

$$F\text{-}Measure =\frac{\mathrm{2RP}}{\mathrm{R+P}} \tag{9}$$

$$G\text{-}Mean =\sqrt{TPR\,.TNR} \tag{10}$$

$$AUC =\frac{1+TPR-FPR}{2} \tag{11}$$

According to [25] in multi-class problems the measurement of classifier performance is determined by an average value. So that the measurement parameters become as follows:

$$Mean\ F-Measure = \sum_{j=1}^{m}\frac{F-Measure(j)}{m} \tag{12}$$

$$Mean\ G-Mean = \sum_{j=1}^{m}\frac{G-Mean(j)}{m} \tag{13}$$

$$Mean\ AUC = \sum_{j=1}^{m}\frac{AUC(j)}{m} \tag{14}$$

$$Mean\ Precision = \sum_{j=1}^{m}\frac{Precision(j)}{m} \tag{15}$$

In (4) it can be seen that the true positive rate (TPR) states the ability of the classifier in classifying a positive sample (minority class) appropriately. the false positive rate (FPR) stated in (5) states the classifier's error in classifying the negative sample (majority class) as a positive (minority class). In (6) states the classifier's ability to correctly classify negative samples. It should be noted that (7) in this case recall is the same as TPR or some other term states as sensitivity. Whereas the Precision stated in (8) is a measure of exactness that states the proportion of positive samples that are classified correctly compared to negative samples that are incorrectly classified as positive samples. The F-Measure and G-Mean stated in (9) and (10) state the ability of the classifier to balance between positive samples accuracy and negative sample accuracy [13]. The AUC stated in (11) states the random probability of a positive sample to be classified correctly compared to the random probability of negative samples [26].

3. METHODOLOGY

The stages of this research can be seen in Figure 1. In Figure 1. the process begins with the determination of the dataset to be used. Sparse Selection and Lasso Penalty calculations are the first processes carried out. The Sparse Selection and Lasso Penalty values obtained will be used for the preprocessing stage by using Feature Selection with minimizing overlapping selection under SMOTE (MOSS). The result of preprocessing with MOSS is in the form of preprocessing dataset which will then enter the processing stage. Processing stages are carried out using different contribution sampling (DCS). The process inside DCS is basically done using the biased support vector machine (B-SVM). Based on the existing classification results, in which this study besides observing the results obtained by the Feature Selection-HAR method, it will also observe the results obtained with MBP + GCE. Then, the result using feature selection-hybrid approach redefinition (HAR) will be compared with the result using MBB + GCE.

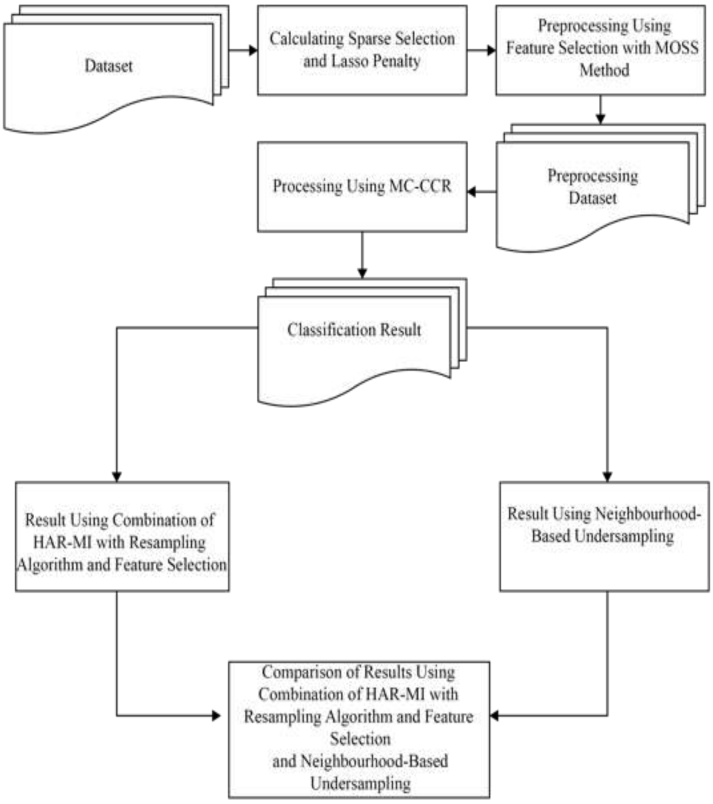
Dataset
Calculating Sparse Selection and Lasso Penalty
Preprocessing Using Feature Selection with MOSS Method
Preprocessing Dataset
Processing Using MC-CCR
Classification Result
Result Using Combination of HAR-MI with Resampling Algorithm and Feature Selection
Result Using Neighbourhood-Based Undersampling
Comparison of Results Using Combination of HAR-MI with Resampling Algorithm and Feature Selection and Neighbourhood-Based Undersampling

Figure 1. Stages of research methods

3.1. Preprocessing stage

Preprocessing Stage on HAR for multi-class problems will be modified using MOSS and using the concept of borderline selection for determining instance in multi-class problems [27]. The preprocessing stages for the proposed method can be seen in the following algorithm.

```
Require: Set S of examples ($x_1$, $y_1$).
Ensure: New set S' of examples with MOSS.
Total size←|S|.
Determine k as the number of Nearest Neighbor.
For All Samples in S do.
Determine the Borderline of Positive or Minority Class as $E_OC_t^+$.
Determine the Borderline of Negative or Majority Class as $E_OC_t^-$.
End For.
For All Samples in $E_OC_t^+$ do.
Calculate the $cn(e)_i$ as neigborhood value for each sample.
Order Ascending the sample according to the $cn(e)_i$.
End For.
```

```
    Building a candidate ensemble for *Safe*, *Borderline*, *Rare*, dan *Outlier* according to $k$ value.
    Calculating Sparse Selection using (2).
    Calculating Loss Penalty using (3).
    Determine sparse regularization.
    For All Samples in *Minority Class* do.
        Sampling All Instance with MOSS.
        Create *newMinority*.
        Create *newMajority*.
    End For.
    Calculate Augmented R-Value.
    Compute the Optimal Loss.
    Return $S^{'}$.
```

In the previous algorithm, it can be seen that after selecting a multi-class dataset, the next step is determining the borderline for minority and majority class. After that, the k value will be determined, then each candidate ensemble will be grouped from the existing instances into Safe, Borderline, Rare, Outlier groups. The next step is to determine the Sparse Selection using (2) and determine the loss penalty using (3). After that for each Instance in minority class, a sampling process using MOSS will be conducted, so that it will produce newminority class and newmajority class. After all these processes are carried out, the Augmented R-Value and optimal loss will be calculated. The results of this preprocessing stage are preprocessing datasets which will then enter the processing stage.

3.2. Processing stage

Processing stage on HAR for multi-class problems will be done using different contribution sampling. Different contribution sampling is an excellent processing method in hybrid ensembles [22]. The processing steps can be seen in the following algorithm.

```
For $i$ = 1 to Number of Instance in Preprocessed Dataset do.
    Add Preprocessed Dataset to $S_{i.}$
    B-SVM will do for classifying $S_{i.}$
    Determine the Majority Class.
    Determine the Minority Class.
    For All Instance in Majority Class do.
        NewSVSets [] will form by checking and delete the noise in SV Sets.
        NewNSVSets [] will form by multiple RUS.
    End while.
    For All instance from new SV Sets and NSV do.
        Create an instance for Majority Class.
    End For.
    For All Instance in Minority Class do.
        SMOTEBoost Process for SV Sets and create SMOTESets.
    end while.
    For All SMOTESets and NewNSVSets do.
        New PositiveSampleSets.
    End For.
    For All NewNegativeSampleSets and NewPositiveSampleSets do.
        ResultDataSet.
    End For.
End For.
```

In the algorithm, it can be seen that, after going through the preprocessing stage, the preprocessing dataset will be obtained, and the next stage of preprocessing this dataset will enter the processing stage. This stage is carried out using DCS through the B-SVM process that will determine instances that exist in the majority and minority class. For each instance the majority class and minority class will then enter the next stage. For each instance of the majority class and minority class will be grouped into support vector sets (SV Sets) and non-support vector sets (NSV Sets) based on the existing hyperplane values. For instances that belong to the SV Sets group in the majority class, it will be checked and removed noise while the NSV Sets will undergo a sampling process using random under sampling (RUS). The results of the RUS process will be combined with the results of noise removal on SV Sets to new majority class. The SV Sets in minority class will undergo the SMOTEBoost process to produce SMOTESets. The SMOTESets and NSV Sets in the Minority class will be combined into a new minority class. New majority class and new minority class will be combined into result dataset.

4. RESULT AND ANALYSIS

4.1. Dataset description

This study uses a multi-class imbalanced dataset that is sourced from the KEEL repository. The dataset selected in this study has represented a low, medium and high imbalance ratio. For datasets with a low imbalance ratio are Hayes-Roth and New-Thyroid, datasets with moderate imbalance ratio are car evaluation and thyroid disease, and dataset with high imbalance ratio are red wine quality, yeast, and shuttle. Dataset description can be seen in Table 2 [28].

Table 2. Dataset description [28]

Dataset	#Ex	#Atts	Distribution of Class	IR
Hayes-Roth	160	4	65/64/31	2.1
New-Thyroid	215	5	150/35/30	5
Car Evaluation	1728	6	384/69/1210/65	18.62
Thyroid Disease	720	21	17/37/666	39.18
Red Wine Quality	1599	11	10/53/681/638/199/18	68.1
Yeast	1484	8	463/5/35/44/51/163/244/429/20/30	92.6
Shuttle	2175	9	1706/2/6/338123	853

4.2. Testing result

The first test is to obtain a comparison of the augmented R-Value and Mean AUC obtained by using feature Selection-HAR and MBP+GCE Method. The test results can be seen in Table 3.

Table 3. Testing result for augmented R-Value and mean AUC for each method

Dataset	Feature Selection-HAR		MBP+GCE	
	Augmented R-Value	Mean AUC	Augmented R-Value	Mean AUC
Hayes-Roth	0.298	0.888	0.295	0.848
New-Thyroid	0.323	0.87	0.331	0.85
Car Evaluation	0.335	0.928	0.336	0.913
Thyroid Disease	0.355	0.871	0.359	0.869
Red Wine Quality	0.412	0.848	0.411	0.797
Yeast	0.432	0.832	0.434	0.799
Shuttle	0.453	0.815	0.461	0.784

Based on the test results, it can be seen that in terms of the Augmented R-Value and also the Mean AUC there is no significant difference between Feature Selection-HAR with MBP + GCE. This shows that both methods have overcome the overlapping problem well. The Augmented R-Value charger is a higher overlap degree of a dataset. For higher imbalance ratio (IR) the results obtained tend to be less good when compared to datasets with lower IR. The second test is to obtain a comparison of the Mean F-Measure, Mean G-Mean, and Mean Precision obtained by using Feature Selection-HAR and MBP+GCE Method. The test results can be seen in Table 4. Based on Table 4. it can be seen that in general the results given by Feature Selection-HAR are better when compared to MBP + GCE.

Table 4. Testing result for mean f-measure, mean g-mean, and mean precision for each method

Dataset	Feature Selection-HAR			MBP+GCE		
	Mean F-Measure	Mean G-Mean	Mean Precision	Mean F-Measure	Mean G-Mean	Mean Precision
Hayes-Roth	0.833	0.884	0.862	0.767	0.841	0.793
New-Thyroid	0.793	0.864	0.821	0.759	0.842	0.786
Car Evaluation	0.634	0.927	0.492	0.626	0.911	0.491
Thyroid Disease	0.789	0.765	0.811	0.763	0.739	0.784
Red Wine Quality	0.609	0.835	0.538	0.5	0.773	0.429
Yeast	0.727	0.816	0.8	0.6	0.774	0.6
Shuttle	0.711	0.797	0.76	0.61	0.749	0.62

4.3. Statistical tests

The statistical test is performed using the Wilcoxon signed-rank test which is a statistical procedure to measure performance based on pairwise comparison [29]. Statistical Tests Result can be seen in Table 5.

Table 5. Wilcoxon signed-rank test for comparing performance measurements

Performance Measurement	P-Value	Hypothesis
Augmented R-Value	0.174749	H_0 (no significant score difference between Feature Selection-HAR and MBP+GCE) is accepted and this means H_1 (there is a significant difference between Feature Selection-HAR and MBP+GCE in score) is rejected because the p-value >0.05
Mean AUC	0.0156250	H_0 (no significant score difference between Feature Selection-HAR and MBP+GCE) rejected and this means H_1 (there is a significant difference between Feature Selection-HAR and MBP+GCE in score) Accepted because the p-value <0.05
Mean F-Measure	0.0156250	H_0 (no significant score difference between Feature Selection-HAR and MBP+GCE) rejected and this means H_1 (there is a significant difference between Feature Selection-HAR and MBP+GCE in score) Accepted because the p-value <0.05
Mean G-Mean	0.0156250	H_0 (no significant score difference between Feature Selection-HAR and MBP+GCE) rejected and this means H_1 (there is a significant difference between Feature Selection-HAR and MBP+GCE in score) Accepted because the p-value <0.05
Mean Precision	0.0156250	H_0 (no significant score difference between Feature Selection-HAR and MBP+GCE) rejected and this means H_1 (there is a significant difference between Feature Selection-HAR and MBP+GCE in score) Accepted because the p-value <0.05

4.4. Discussion

Based on the results in Table 3-5. it can be seen that in general the Feature Selection-HAR method gives results that are not significant difference in the Augmented R-Value between Feature Selection-HAR Method. The results given are relatively good and this means that both methods have handled the overlapping problem well. Based on the value of Mean AUC, F-Measure, G-Mean, and precision, the results given by Feature Selection-HAR give better results compared to MBP + GCE. The statistical test also shows that there are significant differences between the two methods. The difference in results is increasingly visible in datasets with high imbalance ratios, which can provide answers to why the Feature Selection-HAR method is better and at the same time shows that Feature Selection-HAR has a good ability to handle datasets with high imbalance ratios.

It should be noted that the results given by both methods indicate that the IR value does not significantly affect the Mean AUC, F-Measure, G-Mean, and Precision values. The test results show that the number of instances and the number of attributes is very influential on the results of Mean AUC, F-Measure, G-Mean, and Precision. The greater the number of instances and the number of attributes, the results obtained can decrease.

5. CONCLUSION

Based on the test results, it can be seen that in terms of overlapping handling, both methods have been able to obtain satisfactory results, which are indicated by a fairly good Augmented R-Value. Meanwhile, when viewed from the Mean AUC, F-Measure, G-Mean, and Precision values, the results obtained by the Feature Selection-HAR Method are better when compared to the MBP + GCE Method. This is supported by the results of statistical tests using the Wilcoxon signed-rank test. Future Research, must be able to overcome the decline in performance if there is a dataset with a large number of instances and also a large number of attributes. The results showed that both methods have weaknesses when there are datasets with a large number of instances and at the same time also have a large number of attributes.

ACKNOWLEDGEMENTS

This research was funded by a Grant from the Ministry of Education and Culture and the Ministry of Research and Technology of the Republic of Indonesia.

REFERENCES

[1] L. Abdi and S. Hashemi, “To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 238-251, 2016, doi: 10.1109/TKDE.2015.2458858.

[2] J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, “Random Balance: Ensembles of variable priors’ classifiers for imbalanced data,” *Knowledge-Based Systems*, vol. 85, pp. 96-111, 2015, doi: 10.1016/j.knosys.2015.04.022.

[3] A. Fernández, M. J. del Jesus, and F. Herrera, “Addressing Overlapping in Classification with Imbalanced Datasets: A First Multi-objective Approach for Feature and Instance Selection,” in *Intelligent Data Engineering and Automated Learning – IDEAL 2015*, Cham, pp. 36-44, 2015. doi: 10.1007/978-3-319-24834-9_5.

[4] V. García, J. Sánchez, and R. Mollineda, "An Empirical Study of the Behavior of Classifiers on Imbalanced and Overlapped Data Sets," in *Progress in Pattern Recognition, Image Analysis and Applications*, Berlin, Heidelberg, pp. 397-406, 2007, doi: 10.1007/978-3-540-76725-1_42.

[5] G.-H. Fu, Y.-J. Wu, M.-J. Zong, and L.-Z. Yi, "Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics," *Chemometrics and Intelligent Laboratory Systems*, vol. 196, p. 103906, 2020, doi: 10.1016/j.chemolab.2019.103906.

[6] Z. Borsos, C. Lemnaru, and R. Potolea, "Dealing with overlap and imbalance: a new metric and approach," *Pattern Anal Applic*, vol. 21, no. 2, pp. 381-395, 2018, doi: 10.1007/s10044-016-0583-6.

[7] J. Bi and C. Zhang, "An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme," *Knowledge-Based Systems*, vol. 158, pp. 81-93, 2018, doi: 10.1016/j.knosys.2018.05.037.

[8] M. Denil and T. Trappenberg, "Overlap versus Imbalance," in *Advances in Artificial Intelligence*, Berlin, Heidelberg, pp. 220-231, 2010, doi: 10.1007/978-3-642-13059-5_22.

[9] M. Alibeigi, S. Hashemi, and A. Hamzeh, "DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets," *Data & Knowledge Engineering*, vol. 81, pp. 67-103, 2012, doi: 10.1016/j.datak.2012.08.001.

[10] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267-288, 1996.

[11] J. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348-1360, 2001.

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

[13] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A Review," *International Journal of Advances in Soft Computing and Its Application*, vol. 7, no. 3, pp. 176-204, 2015.

[14] E. R. Q. Fernandes and A. C. P. L. F. de Carvalho, "Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning," *Information Sciences*, vol. 494, pp. 141-154, 2019, doi: 10.1016/j.ins.2019.04.052.

[15] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463-484, 2011, doi: 10.1109/TSMCC.2011.2161285.

[16] R. Alejo, R. M. Valdovinos, V. García, and J. H. Pacheco-Sanchez, "A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios," *Pattern Recognition Letters*, vol. 34, no. 4, pp. 380-388, 2013, doi: 10.1016/j.patrec.2012.09.003.

[17] B. Das, N. C. Krishnan, and D. J. Cook, "Handling Class Overlap and Imbalance to Detect Prompt Situations in Smart Homes," in *2013 IEEE 13th International Conference on Data Mining Workshops*, pp. 266-273, 2013, doi: 10.1109/ICDMW.2013.18.

[18] W. Xie, G. Liang, Z. Dong, B. Tan, and B. Zhang, "An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data," *Mathematical Problems in Engineering*, pp. 1-13, 2019, doi: https://doi.org/10.1155/2019/3526539.

[19] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *1*, vol. 61, pp. 863–905, Apr. 2018.

[20] S. Oh, "A new dataset evaluation method based on category overlap," *Comput. Biol. Med.*, vol. 41, no. 2, pp. 115-122, 2011, doi: 10.1016/j.compbiomed.2010.12.006.

[21] K.-W. Hsu, "A Theoretical Analysis of Why Hybrid Ensembles Work," *Computational Intelligence and Neuroscience*, [Online]. Available: https://www.hindawi.com/journals/cin/2017/1930702/. 2017.

[22] H. Hartono, O. S. Sitompul, T. Tulus, E. B. Nababan, and D. Napitupulu, "Hybrid Approach Redefinition (HAR) model for optimizing hybrid ensembles in handling class imbalance: a review and research framework," *MATEC Web Conf.*, vol. 197, p. 03003, 2018, doi: 10.1051/matecconf/201819703003.

[23] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301-320, 2005.

[24] L. Zhang, H. Yang, and Z. Jiang, "Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN," *Biomed Eng Online*, vol. 17, no. 1, p. 181, 2018, doi: 10.1186/s12938-018-0604-3.

[25] R. Alejo, J. A. Antonio, R. M. Valdovinos, and J. H. Pacheco-Sánchez, "Assessments Metrics for Multi-class Imbalance Learning: A Preliminary Study," in *Pattern Recognition*, pp. 335-343, 2013.

[26] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.

[27] S. García, Z.-L. Zhang, A. Altalhi, S. Alshomrani, and F. Herrera, "Dynamic ensemble selection for multi-class imbalanced datasets," *Information Sciences*, vol. 445-446, pp. 22-37, 2018, doi: 10.1016/j.ins.2018.03.002.

[28] J. Alcalá-Fdez *et al.*, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Comput*, vol. 13, no. 3, pp. 307-318, 2009, doi: 10.1007/s00500-008-0323-y.

[29] F. Wilcoxon, "Individual Comparisons by Ranking Methods on JSTOR," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80-83, 1945.

BIOGRAPHIES OF AUTHORS

Hartono obtained his doctoral in Computer Science in 2018 from Universitas Sumatera Utara, completed his master degree in 2010 from the Universitas Putra Indonesia YPTK Padang, Indonesia in Computer Science and bachelor in 2008 from STMIK IBBI Medan, Indonesia in Computer Science. Now he serves as a lecturer at STMIK IBBI and Universitas Potensi Utama in Medan, Indonesia. His current interests are in Machine Learning, Artificial Intelligence, Data Mining and Operational Research.

Erianto Ongko completed his master in 2015 at Universitas Sumatera Utara in Medan, Indonesia in Computer Science and bachelor in 2012 from STMIK IBBI Medan, Indonesia in Computer Science. He is a lecturer at Akademi Teknologi Industri Immanuel in Medan, Indonesia. His current interests are in Machine Learning, Artificial Intelligence and Operational Research.

Yeni Risyani completed a bachelor's degree in computer science from the Institut Sains dan Teknologi TD Pardede in 1997 and a master's degree in computer science from Universitas Putra Indonesia YPTK Padang in 2010. Her current interests are in Artificial Intelligence and Machine Learning.