Clustering-based Ensemble Pruning in the Imbalanced Data Classification

Paweł Zyblewski[0000−0002−4224−6709]

Department of Systems and Computer Networks,
Faculty of Electronics, Wrocław University of Science and Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
pawel.zyblewski@pwr.edu.pl

Abstract. Ensemble methods in combination with data preprocessing techniques are one of the most used approaches to dealing with the problem of imbalanced data classification. At the same time, the literature indicates the potential capability of classifier selection/ensemble pruning methods to deal with imbalance without the use of preprocessing, due to the ability to use expert knowledge of the base models in specific regions of the feature space. The aim of this work is to check whether the use of ensemble pruning algorithms may allow for increasing the ensemble's ability to detect minority class instances at the level comparable to the methods employing oversampling techniques. Two approaches based on the clustering of base models in the diversity space, proposed by the author in previous articles, were evaluated based on the computer experiments conducted on 41 benchmark datasets with a high *Imbalance Ratio*. The obtained results and the performed statistical analysis confirm the potential of employing classifier selection methods for the classification of data with the skewed class distribution.

Keywords: Imbalanced data · Classifier ensemble · Ensemble pruning · Multistage organization

1 Introduction

When dealing with real-life binary classification problems we can often encounter cases, in which the number of samples belonging to one class (also known as majority class) significantly exceeds the number of samples in the other class (called minority class). However, classical pattern recognition algorithms usually assume a balanced distribution of problem instances. Therefore, in the case of skewed class distribution, they tend to display a significant bias towards the majority class. In the literature, three main types of approaches have been distinguished in order to deal with the imbalanced data classification problems [9]:

- *Data-level methods* based on the modification of the training set in such a way as to reduce the bias towards the majority class.
- *Algorithm-level methods* modifying classical pattern recognition algorithms in order to adapt them to deal with imbalance.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

– *Hybrid methods* combining the two above-mentioned approaches.

One of the frequently used approaches to imbalanced data classification is the classifier ensemble [13]. Here, the methods based on the Static and Dynamic Classifier Selection (DCS) are particularly noteworthy [3], as they take into account the base model expertise in specific regions of the feature space. Ksieniewicz in [10] proposed the *Undersampled Majority Class Ensemble* (UMCE), which generates the classifier pool by dividing an unbalanced problem into a series of balanced ones. The *Dynamic Ensemble Selection Decision-making* (DESD) algorithm was presented by Chen et al. [2] in order to employ the weighting mechanism to select base classifiers that are experts in minority class recognition. Wojciechowski and Woźniak [19] employed Decision Templates in order to integrate the classifier pool decisions in case of imbalanced data classification. Klikowski and Woźniak [7] proposed the *Genetic Ensemble Selection* (GES) for imbalanced data, which generated diverse classifier pool based on feature selection using a genetic algorithm.

This article deals with the concept closely related to the classifier selection, knows as ensemble pruning. Zhou in [21] proposed the following taxonomy of such methods:

– *Ranking-based pruning* selecting a certain number of top ranked classifiers, according to a chosen metric [15].
– *Optimization-based pruning* treating the classifier selection problem as an optimization task [18,22].
– *Clustering-based pruning* clustering-based pruning which groups base models making similar decisions, and then selecting prototype classifiers from each cluster to constitute the pruned ensemble.

In this work, the application of clustering-based ensemble pruning algorithms for the imbalanced classification task will be considered. Such methods consist of two steps. First, using the selected clustering algorithm, the base classifiers are grouped in such a way that each cluster contains models that have a similar impact on the ensemble performance. For this purpose, clustering methods such as e.g. *k-means* clustering [4], hierarchical agglomerative clustering [5] and spectral clustering [20] were used. The most important element of clustering-based ensemble pruning methods is defining the space in which clustering takes place. The Euclidean distance was used by Lazarevic and Obradovic [14], while employing the pairwise diversity matrix was proposed by Kuncheva [11].

Then, from each of the clusters, a single model (also knows as the prototype classifier) is selected to be included in the pruned ensemble. For this purpose, e.g., the classifier with the highest accuracy score in [4] or the model farthest from the other clusters [5] can be selected. This step also includes the problem of selecting the number of clusters. It can be determined by evaluating the method on the validation set [4] or, in the case of fuzzy clustering methods, automatically selected using membership values of statistical indexes [8].

The main goal of this work is to examine whether the use of expert classifier knowledge in a given feature space region will allow establishing an ensemble

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

capable of dealing with the imbalanced data classification without the need of using preprocessing techniques.

The main contributions of the following work are as follows:

- Employing proposed Clustering-based Ensemble Pruning methods for the imbalanced data classification problem.
- Experimental evaluation of the proposed algorithms on benchmark datasets and comparison with methods using data preprocessing.

2 Clustering-based pruning and multistage voting organization

This section presents ensemble pruning algorithms based on clustering in the one-dimensional diversity space, which were proposed by Zyblewski and Woźniak in [23,24].

Clustering-based pruning (CPR)
Clustering is performed in the one-dimensional M measure space, which is calculated based on classifier diversity measures. In this work, 5 different diversity metrics were used, namely *the entropy measure* E, *measurement of interrater agreement* k, *averaged disagreement measure* (Dis_{av}), *Kohavi-Wolpert variance* (KW), and the averaged Q statistic (Q_{av}) [12]. The M measure for a given classifier Ψ_i is defined as a difference between the diversity of the whole classifier pool Π and pool without said classifier

$$M(\Psi_i) = Div(\Pi) - Div(\Pi - \Psi_i) \tag{1}$$

An examples of the resulting clustering spaces for each of the diversity measures is shown in Figure 1.

Then the *k-means* clustering algorithm is employed in order to group base classifier with similar effect on the ensemble performance. Finally, from each cluster, a prototype model with the highest *balanced accuracy score* is selected to be a part of the pruned ensemble.

Clustering-based multistage voting organization
Additionally, the *Random Sampling Multistage Organization* (RSMO) algorithm was proposed, that is a modification of the multistage organization, which was first described in [6]. This proposal is based on the *Multistage Organization with Majority Voting* (MOMV) as detailed by Ruta and Gabrys in [17]. This approach can be compared to static classifier selection and ensemble pruning as it selects models for the first voting layer based on sampling with replacement. Sampling usage is based on the assumption that the classifiers within a given cluster have a similar effect on the ensemble performance, therefore they do not have to be all used in the classification process. Then, the final decision is produced by two layers of majority voting. An example of such an organization is shown in Figure 2.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

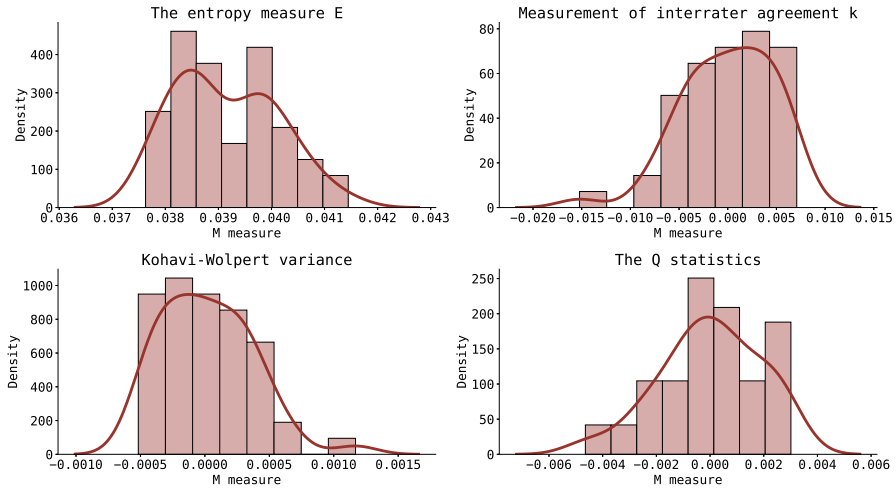

Fig. 1. Histograms and density estimation plots for M measure based on each ensemble diversity metric calculated on the glass2 dataset. *Disagreement measure* was omitted due to the results identical to the *Kohavi-Wolper variance*.

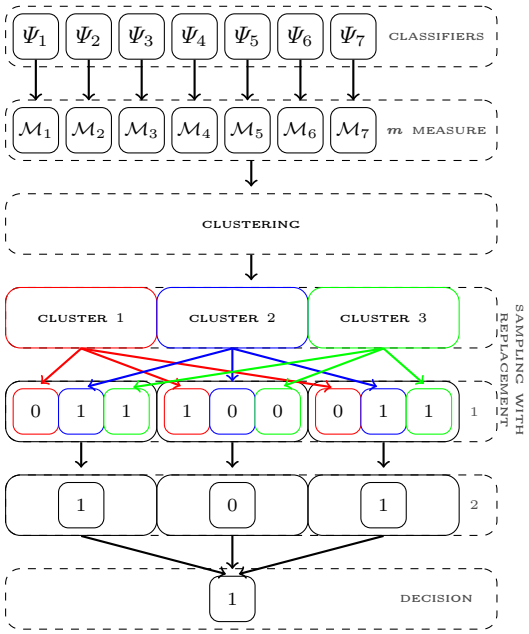

Fig. 2. Example of clustering-based multistage voting organization with 9 classifiers and 3 clusters. The number of groups in the first layer corresponds to the number of clusters. Then, using sampling with replacement, a single classifier from each cluster is selected to be a part of each group.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

Computational and memory complexity analysis
The proposed method includes the stage of determining the M measure value of each base classifier, the clustering of models in the diversity space and the selection of prototype classifiers.

In order to obtain the M measure value for each base classifier, first, the ensemble diversity must be calculated. The complexity of this process is $O(n)$ or $O(n^2)$, where n is the number of base classifiers, depending on whether the non-pairwise or pairwise measure is used. Then, the M measure calculation process has the complexity of $O(n)$.

The *k-means* algorithm was used for clustering in diversity space. Therefore, the complexity of clustering is $O(ncde)$, where c is the number of clusters, d is the number of data dimensions, and e describes the number of iterations/epochs of the algorithm [1]. As the clustering space is one-dimensional, complexity is reduced to $O(nce)$.

3 Experimental evaluation

The research was carried out on 41 imbalanced datasets presented in Table 1, where #I is the number of instances, #F is the number of features and IR denotes the *Imbalance Ratio*. However, it should be noted that the experiments could only be carried out on those datasets for which the *k-means* clustering algorithm was able to find the desired number of clusters (from 2 to 7) for a given classifier and diversity measure.

Table 1. Imbalanced datasets characteristics.

Dataset	#I	#F	IR	Dataset	#I	#F	IR
ecoli-0-1_vs_2-3-5	244	7	9	glass2	214	9	12
ecoli-0-1_vs_5	240	6	11	glass4	214	9	15
ecoli-0-1-3-7_vs_2-6	281	7	39	glass5	214	9	23
ecoli-0-1-4-6_vs_5	280	6	13	led7digit-0-2-4-5-6-7-8-9_vs_1	443	7	11
ecoli-0-1-4-7_vs_2-3-5-6	336	7	11	page-blocks-1-3_vs_4	472	10	16
ecoli-0-1-4-7_vs_5-6	332	6	12	shuttle-c0-vs-c4	1829	9	14
ecoli-0-2-3-4_vs_5	202	7	9	shuttle-c2-vs-c4	129	9	20
ecoli-0-2-6-7_vs_3-5	224	7	9	vowel0	988	13	10
ecoli-0-3-4_vs_5	200	7	9	yeast-0-2-5-6_vs_3-7-8-9	1004	8	9
ecoli-0-3-4-6_vs_5	205	7	9	yeast-0-2-5-7-9_vs_3-6-8	1004	8	9
ecoli-0-3-4-7_vs_5-6	257	7	9	yeast-0-3-5-9_vs_7-8	506	8	9
ecoli-0-4-6_vs_5	203	6	9	yeast-0-5-6-7-9_vs_4	528	8	9
ecoli-0-6-7_vs_3-5	222	7	9	yeast-1_vs_7	459	7	14
ecoli-0-6-7_vs_5	220	6	10	yeast-1-2-8-9_vs_7	947	8	31
ecoli4	336	7	16	yeast-1-4-5-8_vs_7	693	8	22
glass-0-1-4-6_vs_2	205	9	11	yeast-2_vs_4	514	8	9
glass-0-1-5_vs_2	172	9	9	yeast-2_vs_8	482	8	23
glass-0-1-6_vs_2	192	9	10	yeast4	1484	8	28
glass-0-1-6_vs_5	184	9	19	yeast5	1484	8	33
glass-0-4_vs_5	92	9	9	yeast6	1484	8	41
glass-0-6_vs_5	108	9	11				

The evaluation of the proposed methods is based on six metrics used in the case of imbalanced classification problems, i.e. *balanced accuracy score*, *G-mean*, F_1 *score*, *precision*, *recall*, and *specificity*. As base classifiers, *Gaussian Naïve Bayes* Classifier (GNB) and *Classification and Regression Tree* (CART), based on the *scikit-learn* implementation [16], were used. The fixed size of the classifier

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

pool was set to 50 base models, generated using a stratified version of bagging. This bagging generates each bootstrap sampling with replacement majority and minority classes separately while maintaining the original *Imbalance Ratio*. The size of each bootstrap is set to half the size of the original training set. The proposed approaches were evaluated on the basis of 5 times repeated 2-fold cross-validation. The ensemble's decision is based on support accumulation. Statistical analysis of the obtained results was performed using the Wilcoxon rank-sum test ($p = 0.05$). All experiments have been implemented in *Python* programming language and can be repeated using the code on *Github*[1].

Research questions
The conducted research aims to answer two main questions:

1. Is the static classifier selection able to improve the results obtained by combining the entire classifier pool for the task of imbalanced data classification?
2. Can the use of static classifier selection in the problem of imbalanced data classification result in performance comparable with the use of preprocessing techniques?

Goals of the experiments
Experiment 1 – Comparison with standard combination
The aim of the first experiment is to compare the proposed methods with a combination of the entire classifier pool. Support accumulation (SACC) and majority voting (MV) of all 50 base models were used as reference methods. The best of the proposed methods is then used in Experiment 2.

Based on the preliminary study, the following pairs of the *diversity measure:number of clusters* were selected for this experiment:

- CPR GNB – E: 2, k: 2, KW: 2, Dis_{av}: 2, Q_{av}: 3,
- CPR CART – E: 5, k: 3, KW: 3, Dis_{av}: 3, Q_{av}: 5,
- RSMO GNB – E: 6, k: 6, KW: 6, Dis_{av}: 4, Q_{av}: 5,
- RSMO CART – E: 7, k: 7, KW: 7, Dis_{av}: 7, Q_{av}: 3.

Experiment 2 – Comparison with preprocessing techniques
In the second experiment, the methods selected in Experiment 1 are compared with the combination of the whole classifier pool generated using preprocessing methods. Preprocessing is performed separately for each of the bootstraps generated by stratified bagging. *Random Oversampling* (ROS), SMOTE, SVM-SMOTE (SVM) and *Boderline*-SMOTE (B2) were selected as the preprocessing techniques.

3.1 Experiment 1 – Comparison with standard combination

Clustering-based pruning
Figure 3 shows radar plots with the average ranks achieved by each method on all evaluation metrics. For the *gaussian naïve bayes* classifier, the advantage of

[1] `https://github.com/w4k2/iccs21-ensemble-pruning`

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

the proposed methods over the combination of the entire available classifier pool can be observed. The only exception is *recall*, where GNB CPR-E2 is comparable to the reference methods, while the other proposed approaches display a slightly lower average rank value.

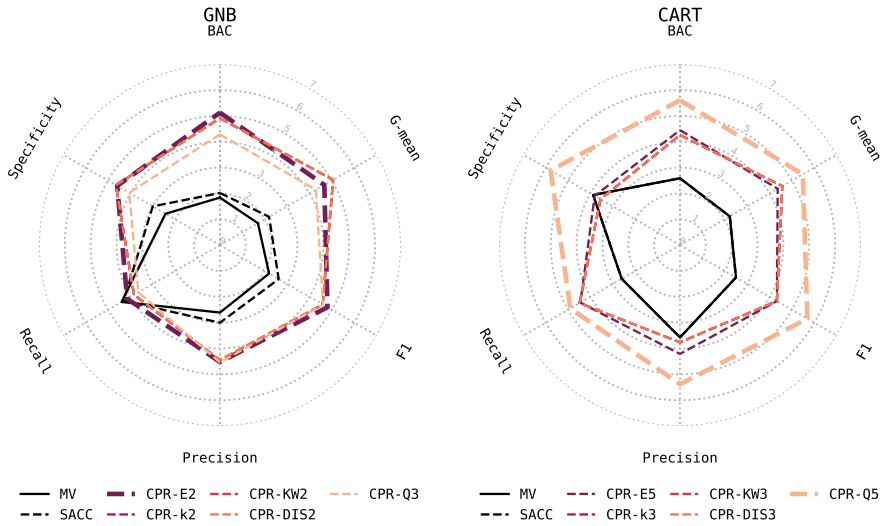

Fig. 3. Visualization of the mean ranks achieved by each method.

These observations are confirmed by Table 2. The numbers under the average rank of each ensemble method indicate which algorithms were statistically significantly worse than the one in question. It presents the results of the performed statistical analysis, on the basis of which it can be concluded that the proposed methods achieve statistically significantly better average ranks than the combination of the entire classifier pool for each of the metrics, except *recall*, where no statistically significant differences were reported. Worth noting is also the identical performance of methods based on measures k, KW, and Dis_{av}.

Particularly promising results can be observed when using CART as the base classifier. In this case, the measure of diversity Q_{av} performs best. Based on the statistical analysis, it achieves statistically significantly better results than the combination of the entire classifier pool, as well as the pruning algorithms using other measures of diversity for the clustering space construction. This is true for every metric except *recall*.

Based on the results of the statistical analysis, the GNB CPR-E2 and CART CPR-Q5 methods were selected for the next experiment. These approaches displayed the highest average ranks as well as a good ability to recognize the minority class.

Random Sampling Two Step Voting Organization

Figure 4 and Table 3 show the comparison of two-step majority voting compared

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

Table 2. Results of Wilcoxon statistical test on global ranks for proposed methods in comparison to the combination of the whole classifier pool.

	GNB						
	MV (1)	SACC (2)	CPR-E2 (3)	CPR-K2 (4)	CPR-KW2 (5)	CPR-DIS2 (6)	CPR-Q3 (7)
BAC	1.839 —	2.018 —	5.125 1, 2	4.911 1, 2	4.911 1, 2	4.911 1, 2	4.286 1, 2
$G-mean$	1.696 —	2.196 1	4.661 1, 2	5.054 1, 2	5.054 1, 2	5.054 1, 2	4.286 1, 2
$F_1 score$	2.196 —	2.625 —	4.804 1, 2	4.589 1, 2	4.589 1, 2	4.589 1, 2	4.607 1, 2
$Precision$	2.607 —	3.000 —	4.518 1, 2	4.446 1, 2	4.446 1, 2	4.446 1, 2	4.536 1, 2
$Recall$	4.393 —	4.304 —	4.143 —	3.839 —	3.839 —	3.839 —	3.643 —
$Specificity$	2.429 —	3.000 1	4.589 1, 2	4.643 1, 2	4.643 1, 2	4.643 1, 2	4.054 1, 2
	CART						
	MV (1)	SACC (2)	CPR-E5 (3)	CPR-K3 (4)	CPR-KW3 (5)	CPR-DIS3 (6)	CPR-Q5 (7)
BAC	2.586 —	2.586 —	4.448 1, 2	4.259 1, 2	4.259 1, 2	4.259 1, 2	5.603 all
$G-mean$	2.224 —	2.224 —	4.362 1, 2	4.569 1, 2	4.569 1, 2	4.569 1, 2	5.483 all
$F_1 score$	2.500 —	2.500 —	4.328 1, 2	4.328 1, 2	4.328 1, 2	4.328 1, 2	5.690 all
$Precision$	3.569 —	3.569 —	4.207 —	3.759 —	3.759 —	3.759 —	5.379 all
$Recall$	2.603 —	2.603 —	4.448 1, 2	4.483 1, 2	4.483 1, 2	4.483 1, 2	4.897 1, 2
$Specificity$	3.879 —	3.879 —	3.810 —	3.552 —	3.552 —	3.552 —	5.776 all

to the reference methods. In the case of GNB RSMO, the most notable is the approach using the Q_{av} diversity measure, which is the most balanced in terms of all evaluation metric. Is is also statistically comparable with the reference methods in terms of the ability to recognize the minority class.

As in the case of GNB, when we use the CART decision tree as the base classifier, the most interesting relationships are represented by the method based on the Q diversity measure. We can see that the CART RSMO-Q3 algorithm achieves highest average ranks in terms of all evaluation metrics. Additionally is is statistically significantly better than reference methods and most of the RSMO approaches using different diversity measures to establish the clustering space.

On the basis of the obtained results, the GNB RSMO-Q5 and CART RSMO-Q3 methods were selected for Experiment 2.

3.2 Experiment 2 – Comparison with preprocessing techniques

Clustering-based pruning
Figure 5 shows the results of comparing the methods selected in Experiment 2 with the approaches employing preprocessing techniques.

When the base classifiers is GNB, it can be noticed that, despite achieving average rank values for each of the metrics, the proposed methods are never statistically significantly worse than the reference approaches using preprocessing (Table 4). Additionally, GNB CPR-E2 shows statistically higher *precision* than that achieved by using Random Oversampling and SMV-SMOTE.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

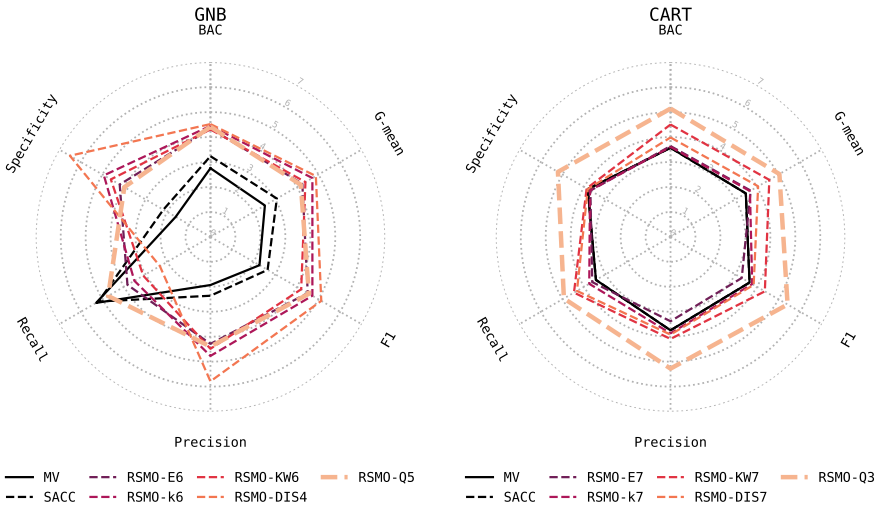

Fig. 4. Average rank values for each of the tested methods.

Table 3. Results of Wilcoxon statistical test on global ranks for proposed methods in comparison to the combination of the whole classifier pool.

	MV (1)	SACC (2)	RSMO-E6 (3)	RSMO-K6 (4)	RSMO-KW6 (5)	RSMO-DIS4 (6)	RSMO-Q5 (7)
GNB							
BAC	2.750 —	3.232 —	4.339 1, 2	4.446 1, 2	4.304 1, 2	4.518 1	4.411 1, 2
$G-mean$	2.518 —	3.071 —	4.250 1, 2	4.714 1, 2	4.393 1, 2	4.875 1, 2	4.179 1, 2
$F_1 score$	2.268 —	2.643 —	4.500 1, 2	4.714 1, 2	4.196 1, 2	5.125 1, 2, 5	4.554 1, 2
$Precision$	1.929 —	2.357 —	4.286 1, 2	4.768 1, 2	4.482 1, 2	5.786 all	4.393 1, 2
$Recall$	5.286 3, 4, 5, 6	5.179 3, 4, 5, 6	3.839 5, 6	3.518 6	3.107 6	2.339 —	4.732 3, 4, 5, 6
$Specificity$	1.607 —	2.179 1	4.196 1, 2	4.929 1, 2, 3, 7	4.607 1, 2	6.500 all	3.982 1, 2
CART							
	MV (1)	SACC (2)	RSMO-E7 (3)	RSMO-K7 (4)	RSMO-KW7 (5)	RSMO-DIS7 (6)	RSMO-Q3 (7)
BAC	3.569 —	3.569 —	3.638 —	3.603 —	4.500 1, 2	3.983 —	5.138 1, 2, 3, 4, 6
$G-mean$	3.483 —	3.483 —	3.707 —	3.672 —	4.569 1, 2	4.052 —	5.034 1, 2, 3, 4, 6
$F_1 score$	3.655 —	3.655 —	3.293 —	3.776 —	4.362 3	3.845 —	5.414 all
$Precision$	3.741 —	3.741 —	3.379 —	3.862 —	4.086 —	3.914 —	5.276 all
$Recall$	3.448 —	3.448 —	3.621 —	3.759 —	4.466 1, 2	4.328 —	4.931 1, 2, 3, 4
$Specificity$	3.828 —	3.828 —	3.707 —	3.724 —	3.879 —	3.828 —	5.207 all

The ensemble pruning methods seem to perform better when using the CART decision tree as the base classifier. Again, none of the reference methods achieved statistically significantly better average ranks than the proposed approach. At

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

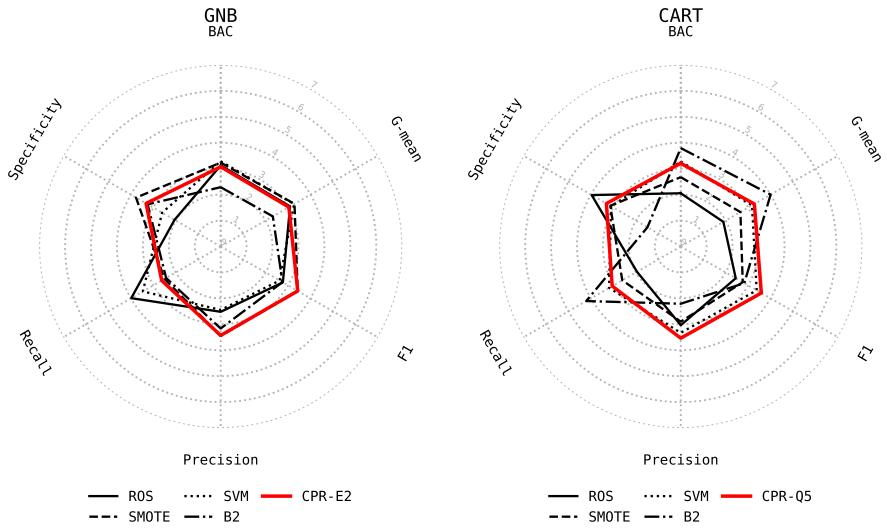

Fig. 5. Visualization of the mean ranks achieved by each method.

Table 4. Results of Wilcoxon statistical test on global ranks for the selected methods in comparison to the preprocessing techniques.

	GNB				
	ROS (1)	SMOTE (2)	SVM (3)	B2 (4)	CPR-E2 (5)
BAC	3.125 4	3.232 4	3.286 4	2.286 —	3.071 —
$G-mean$	3.089 4	3.286 4	3.268 4	2.321 —	3.036 —
$F_1 score$	2.768 —	3.429 3	2.625 —	2.750 —	3.429 —
$Precision$	2.518 —	3.446 1, 3	2.446 —	3.161 —	3.429 1, 3
$Recall$	3.982 2, 4, 5	2.500 —	3.464 2, 4	2.429 —	2.625 —
$Specificity$	2.054 —	3.768 1, 3	2.607 —	3.250 1	3.321 1
	CART				
	ROS (1)	SMOTE (2)	SVM (3)	B2 (4)	CPR-Q5 (5)
BAC	2.052 —	2.672 —	3.276 1, 2	3.793 1, 2	3.207 1
$G-mean$	1.897 —	2.655 1	3.172 1	4.000 1, 2, 3	3.276 1
$F_1 score$	2.448 —	2.759 —	3.379 1, 2	2.828 —	3.586 1, 4
$Precision$	3.034 4	2.897 4	3.328 4	2.207 —	3.534 4
$Recall$	1.948 —	2.603 1	3.190 1, 2	4.207 all	3.052 —
$Specificity$	3.966 2, 3, 4	3.138 4	3.103 4	1.483 —	3.310 4

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

the same time, however, CART CPR-Q5 achieves a statistically significantly better rank value than ROS for BAC, *G-mean* and F_1 *score*. This method is also statistically significantly better than *Borderline*-SMOTE in terms of F_1 *score* and *specificity*.

Random Sampling Two Step Voting Organization
The results of the statistical analysis for the comparison of the proposed RSMO methods with the preprocessing-based approaches shown in Table 5. The average rank values for each of the metrics for are shown in Figure 6.

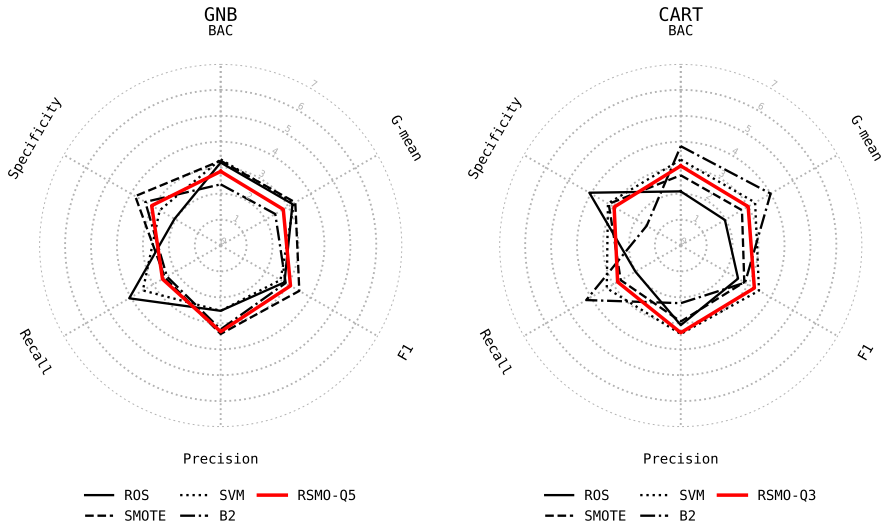

Fig. 6. Average rank values for each of the tested methods

When the base classifier is GNB, the two-step majority voting methods achieve results comparable to *Borderline*-SMOTE, however, they are statistically significantly worse in terms of *recall* than Random Oversampling and SVM-SMOTE. When two-step voting is used in conjunction with the CART decision tree, the proposed method achieves statistically significantly better *precision* than the reference methods. However, it is statistically significantly inferior to *Borderline*-SMOTE in terms of *G-mean* and *recall*.

3.3 Lessons learned

Based on the preliminary experiment determining the appropriate number of clusters for a given diversity measure, it can be concluded that the classifier pool generation using stratified bagging probably does not allow for achieving a high ensemble diversity in the case of GNB. This is indicated by the fact that the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

Table 5. Results of Wilcoxon statistical test on global ranks for the selected methods in comparison to the preprocessing techniques.

	GNB				
	ROS (1)	SMOTE (2)	SVM (3)	B2 (4)	RSMO-Q5 (5)
BAC	3.196 4	3.268 4	3.321 4	2.357 —	2.857 —
$G-mean$	3.196 4	3.321 4	3.268 4	2.429 —	2.786 —
$F_1 score$	2.839 —	3.500 3	2.661 —	2.893 —	3.107 —
$Precision$	2.518 —	3.411 1, 3	2.518 —	3.232 —	3.321 —
$Recall$	4.071 all	2.482 —	3.464 2, 4, 5	2.393 —	2.589 —
$Specificity$	2.054 —	3.804 1, 3	2.714 —	3.357 1	3.071 —
	CART				
	ROS (1)	SMOTE (2)	SVM (3)	B2 (4)	RSMO-Q3 (5)
BAC	2.086 —	2.707 —	3.310 1, 2	3.828 1, 2	3.069 —
$G-mean$	1.966 —	2.724 1	3.310 1	4.000 1, 2, 3	3.000 —
$F_1 score$	2.552 —	2.828 —	3.483 1, 2	2.862 —	3.276 —
$Precision$	3.069 4	2.931 4	3.414 4	2.224 —	3.362 4
$Recall$	2.017 —	2.672 1	3.293 1, 2	4.207 all	2.810 —
$Specificity$	4.069 all	3.207 4	3.241 4	1.517 —	2.966 4

methods using this classifier perform best when the clustering space is divided into just two groups. Decision trees, which show a greater tendency to obtain diverse base models, do much better in this respect. It is also worth noting that in the case of CART, due to no tree depth limitation, the results of the majority vote were in line with the accumulation of support.

Regardless of the base classifier used, the results obtained with the use of the measures of diversity k, KW, and Dis_{av} were exactly the same. On this basis, it can be concluded that the diversity spaces generated on their basis coincide. An example of this can be seen in the example shown in Figure 1, where all three spaces have the same distribution density (where the space based on k is a mirror image of the spaces based on KW and Dis_{av}).

Experiment 1 proved that by a skillful selection of a small group of classifiers, in the imbalanced data classification problem, it is possible to achieve a better performance than that achieved by combining the decisions of the entire classifier pool.

Experiment 2 was able to confirm that thanks to employing the classifier selection methods to the problem of imbalanced data classification, it is possible to obtain results statistically not worse (and sometimes statistically significantly better) than those achieved by the ensembles using preprocessing techniques.

Additionally, from the obtained results, it can be concluded that the use of the two-stage majority voting structure may allow, in the case of imbalanced data classification task, to improve the ensemble performance when compared to the traditional combination of the classifier pool. This is due to the division

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

of classifiers into clusters containing models that make similar errors on problem instances. Thanks to this, after the first voting level, we obtain predictions reflecting the expert knowledge of the base models in each of the recognized feature space regions.

The results of the experiments seem to indicate the averaged Q statistic as the best measure of ensemble diversity for the generation of one-dimensional clustering space. However, according to the research carried out by Kuncheva and Whitaker [12], one cannot indicate the superiority of Q statistics over the other diversity measures.

4 Conclusions

The main purpose of this work was to examine whether the use of static classifier selection/ensemble pruning methods in the imbalanced data classification problems allows for increasing the ensemble's ability to detect minority class instances. The research was conducted using two proposed ensemble pruning methods, based on the base models clustering in one-dimensional diversity space. The obtained results and the performed statistical analysis confirmed that the careful selection of base models in the case of imbalanced data may increase the ability of the pruned ensemble to recognize the minority class. In some cases, such ensembles may even outperform larger classifier pools generated using data oversampling techniques such as *Random oversampling* and *Borderline*-SMOTE.

Future research may include attempts to modify existing classifier selection methods (both static and dynamic) for the purpose of classifying imbalanced data.

Acknowledgment

This work was supported by the *Polish National Science Centre* under the grant No. 2017/27/B/ST6/01325.

References

1. Bora, D.J., Gupta, D., Kumar, A.: A comparative study between fuzzy clustering algorithm and hard clustering algorithm. arXiv preprint arXiv:1404.6059 (2014)
2. Chen, D., Wang, X.J., Wang, B.: A dynamic decision-making method based on ensemble methods for complex unbalanced data. In: Web Information Systems Engineering – WISE 2019. pp. 359–372. Springer International Publishing, Cham (2019)
3. Cruz, R.M., Sabourin, R., Cavalcanti, G.D.: Dynamic classifier selection: Recent advances and perspectives. Inf. Fusion **41**(C), 195–216 (May 2018)
4. Fu, Q., HU, S.X., Zhao, S.: Clustering-based selective neural network ensemble. Journal of Zhejiang University SCIENCE **6**(5), 387–392 (2005)
5. Giacinto, G., Roli, F., Fumera, G.: Design of effective multiple classifier systems by clustering of classifiers. 15th International Conference on Pattern Recognition, ICPR 2000 (2000)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14

6. Ho, T.K., Hull, J.J., Srihari, S.N.: Decision combination in multiple classifier systems. IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(1), 66–75 (Jan 1994)
7. Klikowski, J., Ksieniewicz, P., Woźniak, M.: A genetic-based ensemble learning applied to imbalanced data classification. In: Intelligent Data Engineering and Automated Learning – IDEAL 2019. pp. 340–352. Springer International Publishing, Cham (2019)
8. Krawczyk, B., Cyganek, B.: Selecting locally specialised classifiers for one-class classification ensembles. Pattern Analysis and Applications **20**(2), 427–439 (2017)
9. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence **5**(4), 221–232 (2016)
10. Ksieniewicz, P.: Undersampled majority class ensemble for highly imbalanced binary classification. In: Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications. Proceedings of Machine Learning Research, vol. 94, pp. 82–94. PMLR, ECML-PKDD, Dublin, Ireland (10 Sep 2018)
11. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, Hoboken, NJ (2004)
12. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning **51**(2), 181–207 (2003)
13. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons (2004)
14. Lazarevic, A., Obradovic, Z.: The effective pruning of neural network classifiers. 2001 IEEE/INNS International Conference on Neural Networks, IJCNN 2001 (2001)
15. Margineantu, D.D., Dietterich, T.G.: Pruning adaptive boosting. In: Proceedings of the Fourteenth International Conference on Machine Learning. pp. 211–218. ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)
16. Pedregosa, F., et al: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
17. Ruta, D., Gabrys, B.: A theoretical analysis of the limits of majority voting errors for multiple classifier systems. Pattern Analysis and Applications **2**(4), 333–350 (2002)
18. Ruta, D., Gabrys, B.: Classifier selection for majority voting. Information Fusion **6**(1), 63–81 (2005)
19. Wojciechowski, S., Woźniak, M.: Employing decision templates to imbalanced data classification. In: Hybrid Artificial Intelligent Systems. pp. 120–131. Springer International Publishing, Cham (2020)
20. Zhang, H., Cao, L.: A spectral clustering based ensemble pruning approach. Neurocomputing **139**, 289–297 (2014)
21. Zhou, Z.H.: Ensemble Methods: Foundations and Algorithms. Chapman & Hall CRC, Boca Raton, FL (2012)
22. Zhou, Z.H., Wu, J., Tang, W.: Ensembling neural networks: Many could be better than all. Artif. Intell. **137**(1-2), 239–263 (May 2002)
23. Zyblewski, P., Woźniak, M.: Clustering-based ensemble pruning and multistage organization using diversity. In: International Conference on Hybrid Artificial Intelligence Systems. pp. 287–298. Springer (2019)
24. Zyblewski, P., Woźniak, M.: Novel clustering-based pruning algorithms. Pattern Analysis and Applications pp. 1–10 (2020)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_14