Pattern Analysis and Applications (2020) 23:1049–1058
https://doi.org/10.1007/s10044-020-00867-8

ORIGINAL ARTICLE

Novel clustering-based pruning algorithms

Paweł Zyblewski[1] · Michał Woźniak[1]

Received: 1 August 2019 / Accepted: 3 February 2020 / Published online: 10 February 2020

Abstract
One of the crucial problems of designing a classifier ensemble is the proper choice of the base classifier line-up. Basically, such an ensemble is formed on the basis of individual classifiers, which are trained in such a way to ensure their high diversity or they are chosen on the basis of pruning which reduces the number of predictive models in order to improve efficiency and predictive performance of the ensemble. This work is focusing on clustering-based ensemble pruning, which looks for the group of similar classifiers which are replaced by their representatives. We propose a novel pruning criterion based on well-known diversity measures and describe three algorithms using classifier clustering. The first method selects the model with the best predictive performance from each cluster to form the final ensemble, the second one employs the multistage organization, where instead of removing the classifiers from the ensemble each classifier cluster makes the decision independently, while the third proposition combines multistage organization and sampling with replacement. The proposed approaches were evaluated using 30 datasets with different characteristics. Experimentation results validated through statistical tests confirmed the usefulness of the proposed approaches.

Keywords Ensemble pruning · Classifier ensemble · Clustering · Multistage organization

1 Introduction

Ensemble methods have been a well-known and quickly developing area of research. They owe their success to the fact that their application allows for dealing with a variety of learning problems, such as learning from distributed data sources [23], improving overall classification accuracy [28], learning from data streams [18], hyperspectral image analysis [17] and imbalanced data classification [19]. While in the classic approach only one learner is trained for a given problem, ensemble methods construct many classifiers based on the available training data and combine them to obtain a final decision. The base learners making up the classifier ensemble are trained in such a way that allows for achieving suitable diversity among the classifiers [29]. An ensemble may consist of either heterogeneous or homogeneous models [3]. Heterogeneous classifiers derive, e.g., from employing various learning algorithms to the same training data, while homogeneous classifiers employ different executions of the same learning algorithm (e.g., by differentiating parameters or using different learning set partitions).

Usually, achieving high classification performance by an ensemble is compensated for overall computational complexity growth, because rather than determining the best single classifier, we look for the best-performing set of classifiers and the best combination rule for obtaining the final decision. It is worth mentioning that [13] enumerated two main approaches to design a classifier ensemble, i.e., *coverage optimization*, where the combination rule is given and the main effort is to form an appropriate line-up of individual predictors, and *decision optimization* which aims for finding an optimal combination rule, while the ensemble line-up is fixed.

This work addresses the topic of classifier ensemble pruning, especially clustering-based ensemble pruning methods, in which our goal is to decrease the total number of ensemble members. Due to this, we can improve predictive performance and considerably reduce the computational overhead.

✉ Paweł Zyblewski
pawel.zyblewski@pwr.edu.pl

Michał Woźniak
michal.wozniak@pwr.edu.pl

[1] Department of Systems and Computer Networks, Faculty of Electronics, Wrocław University of Science and Technology, Wybrzeze Wyspianskiego 27, 50-370 Wrocław, Poland

In a nutshell, the main contributions of this work are as follows:

- The proposition of a novel mutual diversity measure based on the non-pairwise and averaged pairwise diversity, which allows to evaluate the impact of a particular predictor on a given classifier ensemble diversity. Thus, it could be used as the criterion for ensemble pruning.
- The formalization of an algorithm that uses the proposed measure for ensemble pruning and multistage organization of majority voting.
- An extensive experimental analysis on a large number of benchmark datasets comparing the performance of proposed methods and the state-of-the-art ensemble methods which are backed up by the statistical tests.

2 Related works

Let us first present the ensemble pruning taxonomy proposed in [32]:

- *Ranking-based pruning* chooses a fixed number of the best-ranked individual classifiers according to a given metric (as kappa statistics) [24].
- *Optimization-based pruning* solves the problem of choosing individual classifiers as an optimization task. Because the number of base models is typically high, therefore heuristic methods [27], evolutionary algorithms [33] or cross-validation-based techniques [5] are usually used.
- *Clustering-based pruning* looks for groups of base classifiers, where individuals in the same group behave similarly while different groups have large diversity. Then, from each cluster, the representative is selected, which is placed in the final ensemble.

Because this work focuses on employing clustering-based classifier ensemble pruning methods to improve the predictive performance of combined classifiers, let us briefly present the main works related to the problem under consideration. Basically, clustering-based pruning consists of two steps. The first one groups base models into several clusters based on a criterion, which should take into consideration their impact on the ensemble performance. For this purpose, various clustering methods were used, such as hierarchical agglomerative clustering [10], deterministic annealing [2], k-means clustering [9, 22] and spectral clustering [31]. Most of those methods employ a kind of diversity-based criteria. Giacinto et al. [10] estimated the probability that classifiers do not make coincident errors in a separate validation set, while Lazarevic and Obradovic [22] used the Euclidean distance in the training set. Kuncheva proposed employing a matrix of pairwise diversity for hierarchical and spectral methods [20].

In the second step, a prototype base learner is selected from each cluster. In [2] a new model was trained for each cluster, based on clusters centroids. In Giacinto et al. [10] choose the classifier, which is the most distant to the rest of clusters. In [22] models were iteratively removed from the least to the most accurate. The model with the best classification accuracy was chosen in [9].

The last issue is the choice of the number of clusters. This could be determined based on the performance of the method on a validation set [9]. In the case of fuzzy clustering methods, we can use indexes based on membership values and dataset or statistical indexes to automatically select the number of clusters [16].

The alternative proposal is a multiple-stage organization, which was briefly mentioned in [14] and described in detail by Ruta and Gabrys [26], where authors refer to such systems as a multistage organization with majority voting (MOMV) since the decision at each level is given by majority voting. Initially, all outputs are allocated to different groups by permutation and majority voting is applied for each group producing single binary outputs, forming the next layer. In the next layers, exactly the same way of grouping and combining is applied with the only difference being that the number of outputs in each layer is reduced to the number of groups formed previously. This repetitive process is continued until the final single decision is obtained. In this research, we employ this approach but to form groups of voting classifier we use clustering methods.

2.1 Ensemble diversity

As mentioned before, diversity is one of the key factors for generating a valuable classifier ensemble, but the main problem is how to measure it. In this work, we decided to use the diversity-based criterion of base classifier clustering. Basically, known diversity measures may be divided into two groups: pairwise and non-pairwise diversity measures. Pairwise diversity measures determine the diversity between pair of base models; ensemble consisting of L classifiers will have $L(L-1)/2$ values of pairwise diversity. To get the value for the entire ensemble, we calculate the average. Non-pairwise measures take into consideration all base classifiers and give one diversity value for the entire ensemble. Let $\varPsi_i$ denote the ith base classifier and $\varPi = \{\varPsi_1, \varPsi_2, \ldots_l\}$ be the ensemble of base models. In this work, three non-pairwise (i.e., entropy measure E, Kohavi–Wolpert variance and measurement of interrater agreement K) and two averaged pairwise (i.e., averaged Q statistics and averaged disagreement measure) ensemble diversity measures have been used. Let us present the selected diversity measures.

The entropy measure E [4] is defined as

$$E(\Pi) = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{L-[L/2]}\right)\min\{l(z_j), L-l(z_j)\}, \tag{1}$$

where N is the number of instances, L stands for the number of base models in the ensemble, and $l(z_j)$ denotes the number of classifiers that correctly recognize z_j. E varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity.

Kohavi–Wolpert variance [15] is defined as

$$\mathrm{KW}(\Pi) = \frac{1}{NL^2}\sum_{j=1}^{N} l(z_j)(L-l(z_j)), \tag{2}$$

The higher the value of KW, the more diverse the classifiers in the ensemble. Also, KW differs from the averaged disagreement measure $\mathrm{Dis}_{\mathrm{av}}$ by a coefficient, i.e.,

$$\mathrm{KW}(\Pi) = \frac{L-1}{2L}\mathrm{Dis}_{\mathrm{av}}(\Pi). \tag{3}$$

Measurement of interrater agreement K [6, 8]

$$K(\Pi) = 1 - \frac{\frac{1}{L}\sum_{j=1}^{N} l(z_j)(L-l(z_j))}{N(L-1)\bar{p}(1-\bar{p})}, \tag{4}$$

where $\bar{p}$ is average individual classification accuracy

$$\bar{p} = \frac{1}{NL}\sum_{j=1}^{N}\sum_{i=1}^{L} y_{j,i}, \tag{5}$$

where $y_{j,i}$ is an element of an N-dimensional binary vector $y_i = [y_{1,i}, \dots, y_{N,i}]^{\mathrm{T}}$ representing the output of a classifier Ψ_i, such that $y_{j,i} = 1$, if Ψ_i recognizes z_j correctly, and 0 otherwise. K varies between 1 and 0, where 1 indicates complete agreement and 0 indicates the highest possible diversity.

The averaged Q statistics [30] over all pairs of classifiers is given as

$$Q_{\mathrm{av}}(\Pi) = \frac{2}{L(L-1)}\sum_{i=1}^{L-1}\sum_{k=i+1}^{L} Q(\Psi_i, \Psi_k), \tag{6}$$

where

$$Q(\Psi_i, \Psi_k) = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \tag{7}$$

and N^{ab} is the number of elements z_j for which $y_{j,i} = a$ and $y_{j,i} = b$. Relationship between a pair of classifiers is denoted according to Table 1. Q varies between −1 and 1. Classifiers that recognize the same objects correctly will have positive values of Q, and those which commit errors on different objects will render Q negative.

The averaged disagreement measure [11] over all pairs of classifiers

Table 1 A table of the relationship between a pair of classifiers

	Ψ_k correct (1)	Ψ_k wrong (0)
Ψ_i correct (1)	N^{11}	N^{10}
Ψ_i wrong (0)	N^{01}	N^{00}

$$\mathrm{Dis}_{\mathrm{av}}(\Psi_i, \Psi_k) = \frac{2}{L(L-1)}\sum_{i=1}^{L-1}\sum_{k=i+1}^{L}\mathrm{Dis}(\Psi_i, \Psi_k), \tag{8}$$

where

$$\mathrm{Dis}(\Psi_i, \Psi_k) = \frac{N^{01} + N^{10}}{N^{11} + N^{10} + N^{01} + N^{00}}. \tag{9}$$

The averaged disagreement measure is the ratio between the number of observations on which one classifier is correct and the other is incorrect to the total number of observations. Dis varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity.

3 Proposed methods

In this section, we propose three methods for increasing the ensemble's accuracy using clustering and diversity-based criterion.

3.1 Clustering criterion

Firstly, let us propose the measure which may be used for the clustering-based pruning. As the non-pairwise and averaged pairwise diversity measures consider all the base models together and calculate one value for the entire ensemble, they could not be used for pruning, because they do not present an impact of a particular base classifier on the ensemble diversity. Therefore, we propose a novel measure M as the clustering criterion, which is the difference between the value of diversity measure for the whole ensemble Π and the value of diversity for the ensemble without a given classifier Ψ_i.

$$M(\Psi_i) = \mathrm{Div}(\Pi) - \mathrm{Div}(\Pi - \Psi_i). \tag{10}$$

Thanks to this proposition, the impact of each base learner on the ensemble diversity is presented in a one-dimensional space, shown in Fig. 1. Each marker represents one of the one hundred base classifiers, placed in the space according to its value of M measure based on the averaged disagreement.

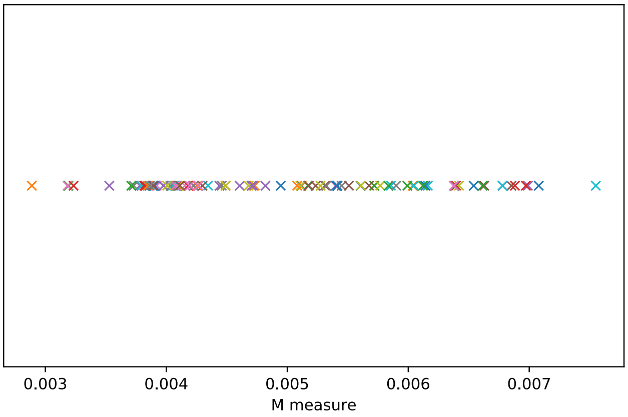

Fig. 1 Visualization of the proposed clustering space for Glass dataset [7], where the clustering criterion (i.e., *M* measure) is calculated based on the entropy measure *E*

3.2 Diversity-based one-dimensional clustering space and cluster pruning

In this proposition, the chosen clustering algorithm is applied to the obtained clustering space. The pruned ensemble consists of the base models with the best classification accuracy in each cluster (one for each cluster).

In case of this work, the *K*-means clustering algorithm, according to the *Scikit-learn* [25] implementation, has been employed to find a given number of clusters (from 2 up to 10) in the clustering space constructed by the proposed *M* measure. From each group, a representative classifier with the highest predictive performance has been chosen. We aim to construct an ensemble containing strong, yet diverse base models, as these two characteristics are distinguishing features of a well-performing classifier ensemble.

3.3 Two-step majority voting organization

The second proposed method is a modification of the MOMV structure described in [26]. Instead of allocating outputs to different groups by permutation, we treat base models in each cluster as a separate ensemble combined by majority voting. Then we collect predictions from each cluster and apply the majority voting rule for the second time, to make a final decision (Fig. 2).

Additionally, we propose the third method, based on the assumption that classifiers belonging to the same cluster make similar decisions, so we do not have to use them all in the classification process. In this method, we construct the first layer of voting by creating the number of groups equal to the number of clusters found, each group containing one classifier sampled with replacement from each of the clusters (Fig. 3).

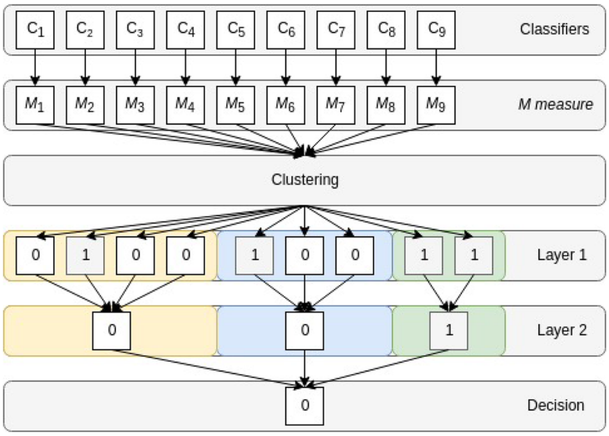
C_1 C_2 C_3 C_4 C_5 C_6 C_7 C_8 C_9
Classifiers
M_1 M_2 M_3 M_4 M_5 M_6 M_7 M_8 M_9
M measure
Clustering
0 1 0 0 1 0 0 1 1
Layer 1
0 0 1
Layer 2
0
Decision

Fig. 2 Example of a two-step majority voting organization with 9 classifiers divided into 3 clusters. Layer 2 is the result of majority voting of each cluster, and the final decision is made by the second majority voting

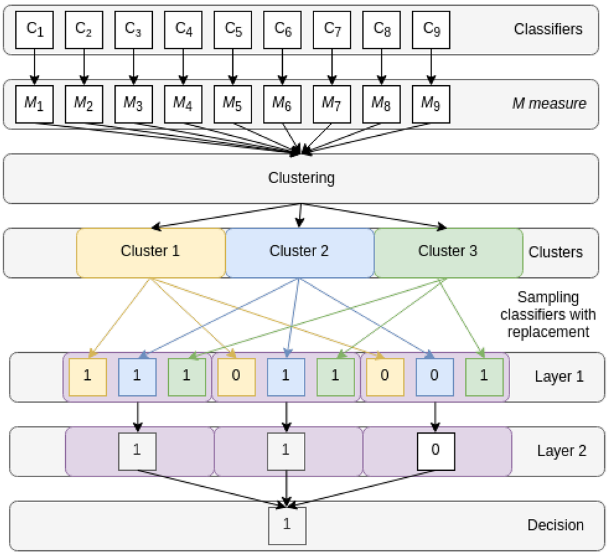
C_1 C_2 C_3 C_4 C_5 C_6 C_7 C_8 C_9
Classifiers
M_1 M_2 M_3 M_4 M_5 M_6 M_7 M_8 M_9
M measure
Clustering
Cluster 1
Cluster 2
Cluster 3
Clusters
Sampling classifiers with replacement
1 1 1 0 1 1 0 0 1
Layer 1
1 1 0
Layer 2
1
Decision

Fig. 3 Example of two-step majority voting organization with 9 classifiers divided into 3 clusters, using sampling with replacement. The number of groups and classifiers in each group in the first layer is equal to the number of clusters found. Layer 2 and the final decision are also made according to the majority voting

4 Experimental study

In this section, we present the experimental study performed to evaluate the effectiveness of the proposed clustering-based ensemble pruning and multistage organization methods. As the reference, two state-of-the-art methods: majority voting and the aggregation of probabilities, were used.

Table 2 Datasets characteristics

Dataset	Instances	Features	Classes	Dataset	Instances	Features	Classes
Appendicitis	106	7	2	NewThyroid	215	5	3
Australian	690	14	2	Pima	768	8	2
Bands	365	19	2	Saheart	462	9	2
Bupa	345	6	2	Segment	2310	19	7
Cleveland	303	13	5	Sonar	208	60	2
Contraceptive	1473	9	3	Spambase	4596	57	2
Dermatology	366	7	8	Spectfheart	267	44	2
Ecoli	336	7	8	Vehicle	846	18	4
Glass	214	9	7	Vowel	990	13	11
Heart	270	13	2	wdbc	596	30	2
HouseVotes	232	16	2	Wine	178	13	3
ILPD	583	10	2	WineRed	1599	11	11
Ionosphere	351	33	2	Winconsin	683	9	2
Libras	360	90	15	Yeast	1484	8	10
MuskV1	476	166	2	ZOO	101	16	7

Experiments were designed to answer the following research questions:

- Which set of parameters (approach, diversity measure, base learner type, number of clusters) yields the best results for the given dataset?
- How the number of clusters affects the performance of methods?
- Does the proposed ensemble pruning and multistage organization methods lead to improvements in accuracy over state-of-the-art methods?

4.1 Datasets

We have used 30 datasets from KEEL [1] and UCI [7] repositories to evaluate the performance of the proposed methods. We have selected a diverse set of benchmarks with varying characteristics, including the different number of instances and features, which are shown in Table 2. Additionally, we take into consideration both binary and multiclass classification problems.

4.2 Setup

As base learners, we used four popular types of classifiers: multilayer perceptron (MLP), classification and regression trees (CART), Gaussian naïve Bayes (NB) and k-nearest neighbors classifier (KNN). In each case, learners from Scikit-learn machine learning library [25] with the default parameters were used. The classifier pool always consists of 100 base models. Diversity between learners is based on the *random subspace method* [12], where classifiers are trained on pseudorandomly selected subsets of components of the feature vector. The percentage of features for training a single model has been selected depending on the number of features in the dataset. For majority of datasets it is 50%, the only exceptions being: Libras dataset—20%, MuskV1 dataset—10%, Sonar dataset—25%, Spambase dataset—25% and Spectfheart dataset—35%. We change the percentage of features used for training so that, regardless of their total number in a given dataset, only a maximum of a dozen or so features were used to train each of the base models to ensure the high diversity.

Based on 3 parameters (approach, diversity measure and base learner type), we distinguish 60 different methods for improving classification score of the ensemble (20 for pruning, 20 for multistage organization and 20 for MO using sampling with replacement). Experiments were carried out for the number of clusters in the range from 2 to 10. For the sake of simplicity, for each method, we take into account only the number of clusters that obtained the best classification accuracy. The name of each method is based on abbreviations of parameter values (*Approach-ClassifierDiversityMeasure* format) including two state-of-the-art methods (majority voting and aggregation of probabilities) for each base learner, which gives us 68 methods overall. The following abbreviations have been used:

- *Approach* MV—majority voting, Aggr—aggregation of probabilities, Mo—multistage voting organization, MoR—multistage voting using sampling with replacement and Pr—clustering-based pruning,
- *Classifier* Mlp—Multilayer perceptron, Cart—classification and regression trees, Nb—Gaussian naïve Bayes and Knn—k-nearest neighbors classifier,
- *DiversityMeasure* E—the entropy measure, KW—Kohavi–Wolpert variance, K—measurement of interrater

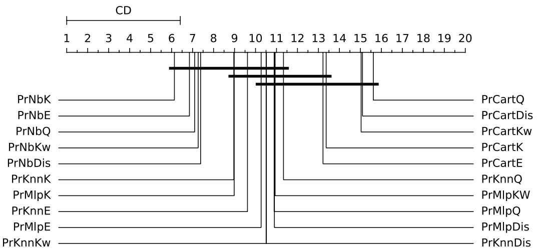

Fig. 4 Diagram of critical difference (CD) for Nemenyi post hoc test at $\alpha = 0.05$ for Pr methods. CD = 5.41

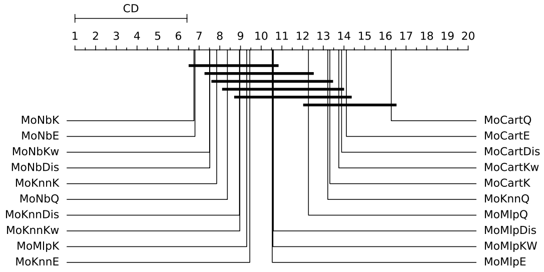

Fig. 5 Diagram of critical difference (CD) for Nemenyi post hoc test at $\alpha = 0.05$ for Mo methods. CD = 5.41

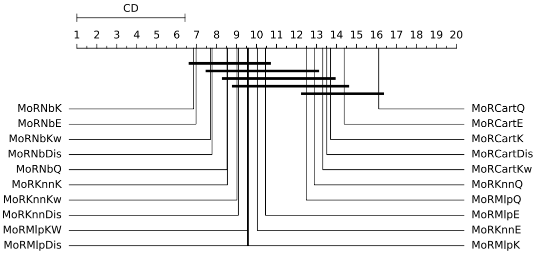

Fig. 6 Diagram of critical difference (CD) for Nemenyi post hoc test at $\alpha = 0.05$ for MoR methods. CD = 5.41

agreement, Q—the averaged Q statistics and Dis—the averaged disagreement measure.

Experiments were implemented in Python programming language and may be repeated according to source code published on *GitHub*.[1]

4.3 Statistical evaluation

First, the proposed methods were divided into 3 groups of 20, based on the used approach. For each group, Nemenyi post hoc test, based on the average ranks according to classification score, was performed (Figs. 4, 5 and 6). In each case, methods employing classification and regression trees as base models achieved the highest average ranks, while methods using Gaussian naïve Bayes classifiers performed the worst.

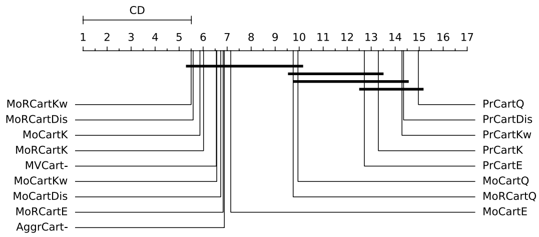

Fig. 7 Diagram of critical difference (CD) for Nemenyi post hoc test at $\alpha = 0.05$ for CART methods. CD = 4.51

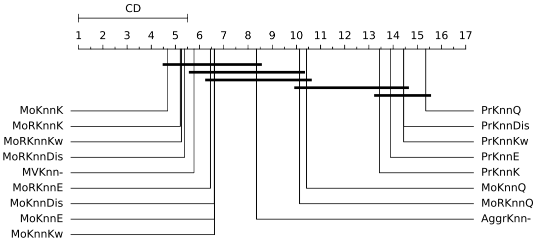

Fig. 8 Diagram of critical difference (CD) for Nemenyi post hoc test at $\alpha = 0.05$ for KNN methods. CD = 4.51

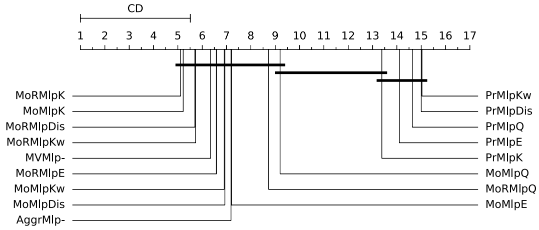

Fig. 9 Diagram of critical difference (CD) for Nemenyi post hoc test at $\alpha = 0.05$ for MLP methods. CD = 4.51

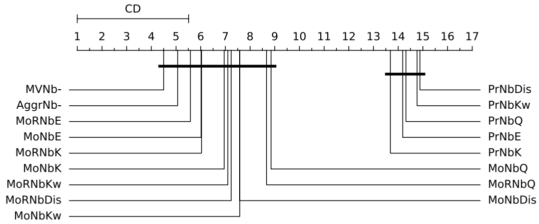

Fig. 10 Diagram of critical difference (CD) for Nemenyi post hoc test at $\alpha = 0.05$ for NB methods. CD = 4.51

Figures 7 8, 9 and 10 show CD diagrams for the proposed methods depending on the type of base models used. In Fig. 7, we can see that, among CART methods, pruning

[1] https://github.com/w4k2/PAA-clustering-based-pruning.

Table 3 The classification accuracy of the best-performing method for each dataset, depending on the number of clusters

Dataset	BestMethod	2C	3C	4C	5C	6C	7C	8C	9C	10C
Appendicitis	PrCartQ	90.56	*91.52*	90.56	89.65	89.61	88.66	90.56	89.61	89.61
Australian	PrCartQ	82.47	88.41	86.24	89.57	88.7	90.73	89.58	*90.88*	90.15
Bands	PrCartK	69.86	76.71	76.99	82.19	80.27	81.37	82.47	*84.11*	82.74
Bupa	PrMlpE	70.43	73.91	74.78	72.75	74.78	74.78	73.91	*76.23*	74.2
Cleveland	PrCartKw	62.28	62.95	62.58	62.95	63.62	64.3	*66.34*	63.31	62.32
Contraceptive	PrMlpQ	55.67	56.89	56.21	56.28	56.08	56.21	56.83	*57.71*	56.62
Dermatology	PrMlpKw	96.4	98.33	97.52	99.72	98.87	99.72	*100.0*	99.17	99.44
Ecoli	PrMlpE	82.5	84.31	84.82	86.35	86.06	86.37	86.96	87.81	*87.84*
Glass	PrCartK	74.72	84.07	80.98	81.3	81.73	84.13	85.43	*86.85*	85.95
Heart	PrCartKw	77.78	87.04	82.96	87.41	84.81	*90.0*	87.78	86.67	87.04
HouseVotes	PrMlpQ	*97.86*	96.58	97.44	96.56	96.58	94.83	96.14	93.98	94.85
ILPD	PrCartK	*76.49*	74.61	75.12	74.1	73.41	74.78	74.1	74.27	74.27
Ionosphere	PrCartDis	93.16	95.73	97.44	98.01	98.01	97.15	98.29	97.73	*98.58*
Libras	PrCartK	72.87	76.2	80.8	83.33	85.4	84.07	86.33	85.87	*87.13*
MuskV1	PrMlpKw	86.57	89.48	90.34	93.91	92.44	96.02	95.18	95.81	*96.44*
NewThyroid	PrNbKw	94.42	*97.21*	94.42	96.28	95.81	96.74	94.88	96.74	–
Pima	PrNbK	74.87	78.39	75.26	*79.95*	76.04	76.95	75.91	76.17	76.3
Saheart	PrNbE	75.31	76.61	75.96	*77.69*	77.26	75.96	76.39	75.32	75.09
Segment	PrCartQ	96.71	97.92	98.27	98.48	98.53	98.61	98.61	98.7	*98.74*
Sonar	PrCartQ	85.59	88.48	88.45	92.33	90.83	*95.7*	92.77	94.71	93.75
Spambase	PrCartQ	89.45	93.08	92.73	93.82	93.65	94.56	94.41	*94.98*	94.39
Spectfheart	PrCartQ	81.62	86.49	87.99	87.99	88.37	89.13	*89.5*	88.36	87.99
Vehicle	PrCartK	72.22	77.31	78.26	80.98	80.27	81.21	81.33	82.04	*82.05*
Vowel	PrKnnE	87.58	91.92	94.14	94.65	95.35	95.66	95.35	95.86	*96.57*
wdbc	PrCartE	95.43	98.07	96.84	97.37	97.72	97.9	97.72	*98.07*	–
Wine	PrCartKw	96.11	96.08	99.46	98.87	98.32	99.43	*100.0*	99.43	*100.0*
WineRed	PrCartK	66.36	67.1	69.41	68.98	69.48	69.61	70.79	70.23	*71.11*
Wisconsin	PrCartQ	97.37	98.1	98.68	98.54	98.68	98.39	98.68	98.68	*98.83*
Yeast	PrMlpE	48.57	55.98	51.68	56.66	53.56	*57.47*	56.12	57.4	55.52
ZOO	PrCartKw	96.99	99.05	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*

The highest achieved values of the classification accuracy have been marked

approaches achieved the highest average ranks and are statistically significantly better than the state-of-the-art and most multistage organization methods. It is worth mentioning the fact that the best-ranked method for every tested approach used the averaged Q statistics as the diversity measure for constructing the clustering space. The same is true for KNN methods (Fig. 8).

In the case of MLP (Fig. 9) and NB (Fig. 10), we can see that methods employing pruning are statistically significantly better than the rest of the proposed and state-of-the-art approaches. Also, the averaged Q statistics again has been the best diversity method for constructing the clustering space, when used for multistage organization methods.

Table 3 presents the impact of the number of clusters on the best-performing methods, according to the classification score, for each tested dataset. As it was not possible to find ten clusters for each method and dataset, we present the maximum number of clusters found in each of the k-folds during evaluation. In the case where several methods achieved the same classification accuracy, the first one was chosen according to the order: *Aggr-*, *MV-*, *Pr*(*E/Kw/K/Dis/Q*), *Mo*(*E/Kw/K/Dis/Q*), *MoR*(*E/Kw/K/Dis/Q*). For every dataset, the proposed pruning methods achieved the best classification accuracy. Figures 11, 12 and 13 present how the performance of methods varies depending on the number of clusters.

4.4 Lessons learned

Increasing the number of clusters positively impacts the classifier performance for the majority of tested classifiers, yet, sometimes we observe a decrease in the classification accuracy after exceeding a certain number of clusters, which may be caused by overfitting leading to a non-optimal number of clusters for a given problem.

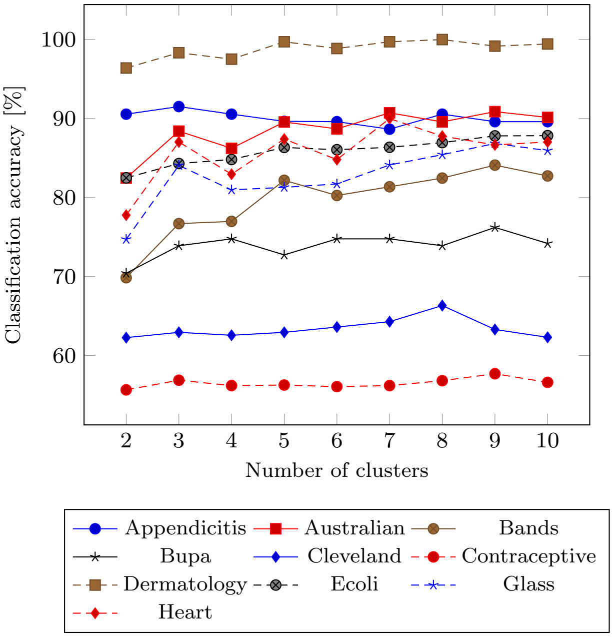

Fig. 11 The classification accuracy of the best-performing methods for different numbers of clusters

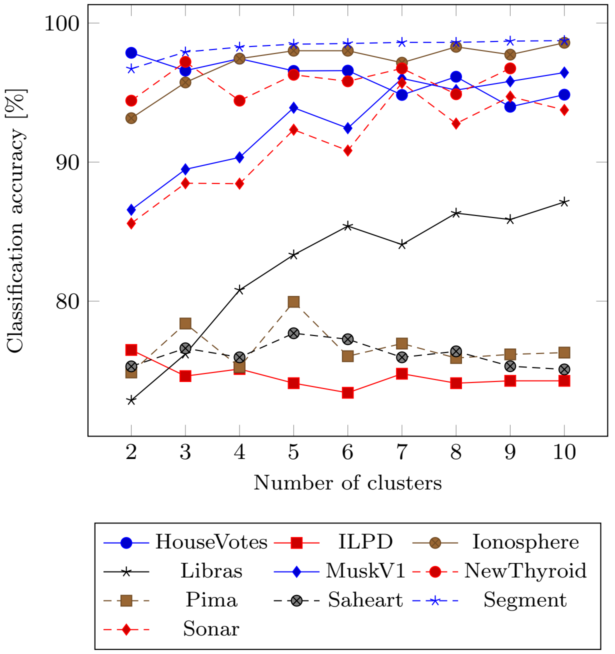

Fig. 12 The classification accuracy of the best-performing methods for different numbers of clusters

The low classification score in the case of two clusters may be caused by using only two classifiers for majority voting. In that case, when there is no agreement between classifiers, the first label in the order is chosen. In the case

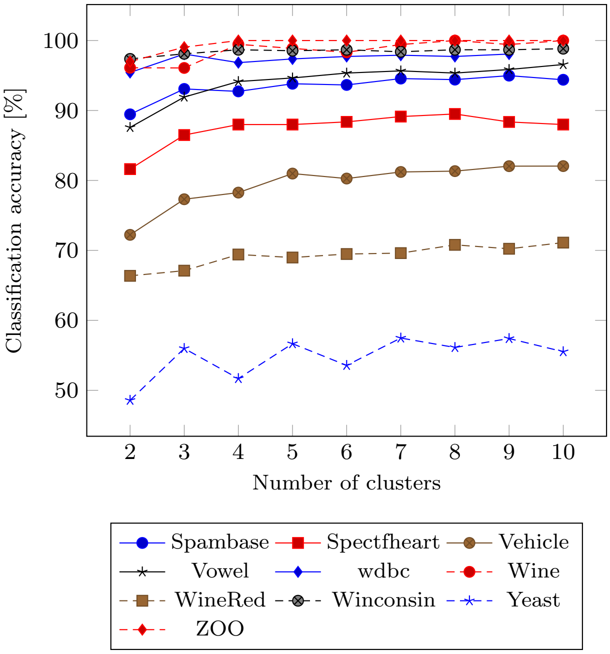

Fig. 13 The classification accuracy of the best-performing methods for different numbers of clusters

of several datasets (e.g., Australian, Contraceptive, Sonar or Yeast), we can observe the reduction in the quality of classification when there is an even number of base models in the ensemble. These reductions may be the result of voting ties and also selecting the first available label as the final decision.

In the case of some datasets (i.e., Dermatology, Wine, and ZOO), we may see that increasing the number of clusters for the proposed pruning method (and thus creating larger ensembles) resulted in achieving a 100% classification accuracy. We may conclude that the proposed clustering-based method really allows for choosing suitably diverse base models, which can create a well-complementing and strong classifier ensemble.

We can also notice the trend for proposed algorithms to perform better on higher-dimensional datasets (e.g., Ionosphere, Libras, MuskV1 or Spambase) when the higher number of clusters is discovered. Similar observation does occur in the case of datasets with more possible class labels (e.g., Libras, Vowel or WineRed).

Although conducted statistical tests indicate that the most suitable diversity measure for the problems considered during experimentation may be the averaged Q statistics, we cannot definitively consider it the best. As stated in [21], after studying various diversity measures, there is no definitive connection between the measures and the improvement of the accuracy and Q_{av} was recommended only based on ease of interpretation and calculation.

5 Conclusions

The main aim of this work was to propose a novel, effective classifier pruning method based on clustering. We proposed the one-dimensional clustering space based on ensemble diversity measures, which is later used in order to prune the existing classifier pool or to perform a multistage majority voting. The computer experiments confirmed the usefulness of the proposed pruning method and based on a statistical analysis we may conclude that it is statistically significantly better than state-of-the-art ensemble methods. It is also worth noting that the pruning approach performed the best among the three methods proposed in this paper. The proposed multistage organization voting scheme (both using the whole classifier pool and sampling with replacement) did not achieve statistically better results than state-of-the-art methods.

The results presented in this paper are quite promising; therefore, they encourage us to continue our work on employing clustering-based methods for ensemble pruning. Future research directions may include exploring the different ways of calculating the proposed *M* measure (including both deterministic and non-deterministic variants) and, in the case of multistage organization methods, employing different types of voting (e.g., weighted majority voting). It would be useful to also consider ways of dealing with ties during the voting process and, possibly, investigate the effects of data dimensionality on the performance of the proposed algorithms.

Acknowledgements This work was supported by the Polish National Science Centre under the Grant No. 2017/27/B/ST6/01325 as well as by the statutory funds of the Department of Systems and Computer Networks, Faculty of Electronics, Wroclaw University of Science and Technology.

References

1. Alcalá-Fdez J, Fernández A, Luengo J, Derrac J, García S (2011) Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. Mult Valued Log Soft Comput 17(2–3):255–287
2. Bakker B, Heskes T (2003) Clustering ensembles of neural network models. Neural Netw 16(2):261–269
3. Bian S, Wang W (2007) On diversity and accuracy of homogeneous and heterogeneous ensembles. Int J Hybrid Intell Syst 4(2):103–128
4. Cunningham P, Carney J (2000) Diversity versus quality in classification ensembles based on feature selection. In: López de Mántaras R, Plaza E (eds) Machine learning: ECML 2000. Springer, Berlin, Heidelberg, pp 109–116
5. Dai Q (2013) A competitive ensemble pruning approach based on cross-validation technique. Knowl Based Syst 37:394–414
6. Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 40(2):139–157
7. Dua D, Graff C (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml. Accessed 23 May 2019
8. Fleiss JL (1981) Statistical methods for rates and proportions. Wiley, Hoboken
9. Fu Q, SX HU, Zhao S (2005) Clustering-based selective neural network ensemble. J Zhejiang Univ Sci 6(5):387–392
10. Giacinto G, Roli F, Fumera G (2000) Design of effective multiple classifier systems by clustering of classifiers. In: 15th International conference on pattern recognition, ICPR 2000
11. Ho TK (1998a) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844
12. Ho TK (1998b) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20:832–844
13. Ho TK (2000) Complexity of classification problems and comparative advantages of combined classifiers. In: Multiple classifier systems. Springer, Berlin Heidelberg, pp 97–106
14. Ho TK, Hull JJ, Srihari SN (1994) Decision combination in multiple classifier systems. IEEE Trans Pattern Anal Mach Intell 16(1):66–75
15. Kohavi R, Wolpert D (1996) Bias plus variance decomposition for zero-one loss functions. In: Proceedings of the thirteenth international conference on international conference on machine learning. Morgan Kaufmann Publishers Inc., San Francisco, ICML'96, pp 275–283
16. Krawczyk B, Cyganek B (2017) Selecting locally specialised classifiers for one-class classification ensembles. Pattern Anal Appl 20(2):427–439
17. Krawczyk B, Ksieniewicz P, Woźniak M (2014) Hyperspectral image analysis based on color channels and ensemble classifier. In: Pan JS, Woźniak M, Quintian H, Corchado E, Polycarpou M, de Carvalho ACPLF (eds) Hybrid artificial intelligence systems. Springer, Cham, pp 274–284
18. Krawczyk B, Minku LL, Gama J, Stefanowski J, Wozniak M (2017) Ensemble learning for data stream analysis: a survey. Inf Fusion 37:132–156
19. Ksieniewicz P (2019) Combining random subspace approach with smote oversampling for imbalanced data classification. In: Pérez García H, Sánchez González L, Castejón Limas M, Quintián Pardo H, Corchado Rodríguez E (eds) Hybrid artificial intelligent systems. Springer, Cham, pp 660–673
20. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, Hoboken

21. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51(2):181–207
22. Lazarevic A, Obradovic Z (2001) The effective pruning of neural network classifiers. In: 2001 IEEE/INNS international conference on neural networks, IJCNN 2001
23. Li Y, Bai C, Reddy CK (2016) A distributed ensemble approach for mining healthcare data under privacy constraints. Inf Sci 330:245–259
24. Margineantu DD, Dietterich TG (1997) Pruning adaptive boosting. In: Proceedings of the fourteenth international conference on machine learning. Morgan Kaufmann Publishers Inc., San Francisco, ICML '97, pp 211–218
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
26. Ruta D, Gabrys B (2002) A theoretical analysis of the limits of majority voting errors for multiple classifier systems. Pattern Anal Appl 2(4):333–350
27. Ruta D, Gabrys B (2005) Classifier selection for majority voting. Inf Fusion 6(1):63–81
28. Wozniak M (2013) Hybrid classifiers: methods of data, knowledge, and classifier combination, vol 519. Springer, Berlin
29. Woźniak M, Graña M (2014) A survey of multiple classifier systems as hybrid systems. Inf Fusion 16:3–17
30. Yule GU (1900) On the association of attributes in statistics. Philos Trans A(194):257–319
31. Zhang H, Cao L (2014) A spectral clustering based ensemble pruning approach. Neurocomputing 139:289–297
32. Zhou ZH (2012) Ensemble methods: foundations and algorithms. Chapman & Hall CRC, Boca Raton
33. Zhou ZH, Wu J, Tang W (2002) Ensembling neural networks: many could be better than all. Artif Intell 137(1–2):239–263

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.