Neurocomputing 160 (2015) 120–131

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Relating ensemble diversity and performance: A study in class noise detection

Borut Sluban [a,*], Nada Lavrač [a,b]

[a] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

ARTICLE INFO

Article history:
Received 15 March 2014
Received in revised form
10 October 2014
Accepted 11 October 2014
Available online 11 February 2015

Keywords:
Class noise
Label noise
Noise detection
Ensemble methods
Diversity measures

ABSTRACT

The advantage of ensemble methods over single methods is their ability to correct the errors of individual ensemble members and thereby improve the overall ensemble performance. This paper explores the relation between ensemble diversity and noise detection performance in the context of ensemble-based class noise detection by studying different diversity measures on a range of heterogeneous noise detection ensembles. In the empirical analysis the majority and the consensus ensemble voting schemes are studied. It is shown that increased diversity of ensembles using the majority voting scheme does not lead to better noise detection performance and may even degrade the performance of heterogeneous noise detection ensembles. On the other hand, for consensus-based noise detection ensembles the results show that more diverse ensembles achieve higher precision of class noise detection, whereas less diverse ensembles lead to higher recall of noise detection and higher *F*-scores.

1. Introduction

In data mining, the success of learning and knowledge discovery from the data depends on various factors, including data quality. The quality of real-life data is frequently degraded due to errors and other data irregularities that are usually referred to as *noise*. The presence of noise has adverse effects on the quality of information retrieved from the data, models created from the data and decisions made based on the data [1]. Given that identifying noisy instances in the data and removing or correcting them proved to be beneficial in various applications, noise identification and filtering became an established area of machine learning and data mining research [2].

Noise in the data manifests itself as attribute noise (errors or unusual attribute values), class noise (wrong instance labels), or a combination of both. Noise detection algorithms are designed to identify erroneous data instances, which are typically found as those deviating from the expected distribution or not following a general pattern or model describing the data. Since every noise detection approach may perform best on a certain domain or on a certain type of noise, the overall noise detection performance can be improved by using ensembles of noise detection algorithms.

Ensemble learning methods are algorithms that construct a set of prediction models (an ensemble) and combine their outputs to a single prediction [3]. Ensembles are typically used with the purpose of improving the performance of simple base learning methods. The strength of ensemble methods lies in their ability to correct errors made by some of their members [4]. Therefore, ensemble members have to be diverse in terms of the errors they make, so that their combination can reduce the total prediction error [5]. Ensembles with greater diversity among their members tend to result in higher predictive accuracy [6].

Diversity among the members of an ensemble can be achieved in different ways, resulting in homogeneous or heterogeneous ensembles. On one hand, in homogeneous ensembles all ensemble members use the same learning algorithm. Popular methods based on boosting [7] and bagging [8], which construct homogeneous ensembles, diversify ensemble members by training them on differently selected subsets of the training data. Some approaches prefer to use different parameter settings of algorithms in the training phase to obtain different classifiers. Other approaches, like the Random Forest algorithm [9], use different feature subsets for training the base classifiers. On the other hand, heterogeneous ensembles are constructed from different base algorithms. It was shown that heterogeneous ensembles are more diverse [10] and that they provide better results than homogeneous ensembles [11]. Heterogeneous ensembles can be constructed by *ensemble selection* [12–14] or *ensemble pruning* [15,16], or be used for meta-learning called *stacking* [17,18]. Ensemble selection and ensemble

* Corresponding author.
E-mail address: borut.sluban@ijs.si (B. Sluban).

http://dx.doi.org/10.1016/j.neucom.2014.10.086
0925-2312/

pruning try to select the base classifiers by balancing the diversity and the performance of the ensemble, while stacking constructs a higher-level predictive model based on the predictions of the first-level base models.

Various measures for assessing the diversity of classifiers have been proposed in the literature [19,5,20,16,21]. The influence of diversity on ensemble performance has been extensively explored for classification problems by observing classification accuracy and classification error rates [5,22–25]. Some studies observed a positive correlation between diversity and classification accuracy [6,26], whereas others doubted that diversity measures can be used as means for improving classification performance [22,27].

In contrast with the above studies of the effects of ensemble diversity on classification accuracy, this paper focuses on the effects of ensemble diversity on the performance of explicit noise detection, which can be used for data cleaning, improved data understanding, and semi-supervised outlier identification, as studied in [28–32]. In these tasks, the main goal is to achieve high performance of explicit noise detection, rather than to increase the classification accuracy of learning algorithms applied after the noise filtering step. In the paper we explore the relation between different diversity measures and the performance of explicit noise detection, achieved by heterogeneous noise detection ensembles.

To the best of our knowledge, this is the first study that directly addresses the relation between different diversity measures and the performance of heterogeneous noise detection ensembles. Note that ensemble-based approaches to noise detection found in the literature recognize the diversity among ensemble members as a requirement for good ensemble performance, however they cope with ensemble diversity only indirectly. Commonly a heterogeneous set of presumably diverse approaches to noise detection is used [32–36], or the diversity of noise identification models is achieved by sampling of the training space and by random selection of features [37–41], or a combination of both approaches is adopted [42,43]. The reason for not explicitly measuring ensemble diversity may lie in the absence of a uniformly accepted definition of diversity. To fill this void, this work studies the relation between different commonly used diversity measures and the performance of various noise detection ensembles. In further work, these results can be used as guidance in the construction of noise detection ensembles.

The rest of the paper is structured as follows. Section 2 introduces the noise detection algorithms, the performance measures used in the evaluation of explicit noise detection, and the measures used for measuring the diversity of ensembles of noise detection algorithms. In Section 3 the aim of the paper is further clarified by presenting the research hypothesis and the goals, followed by the proposed methodology and experimental setting used in evaluating the relation between ensemble diversity and noise detection performance. The experimental results are presented in Section 4. The paper concludes in Section 5 with a discussion of the obtained results and directions for further work.

2. Preliminaries

This section introduces the basic methods and measures required for studying the relation between ensemble diversity and noise detection performance. First, class noise detection is described, second the performance measures for noise detection evaluation are specified, and finally, a selection of commonly used ensemble diversity measures is presented.

2.1. Noise detection

Class noise denotes errors in the labels assigned to data instances. From a wide variety of noise handling techniques [2], we chose a popular class noise detection approach proposed in [33], which became to be later known as *classification noise filtering*. This approach uses classification algorithms to identify wrongly labeled data instances. It works in a k-fold cross-validation manner, where in k repetitions $k-1$ folds of the dataset are used for training of a classification algorithm and the complementary fold is used for classifier validation. The instances that are misclassified on the validation folds are identified as noisy. The concept of classification noise filtering is illustrated in Fig. 1.

In the experiments we will investigate the performance of heterogeneous ensembles of classification noise filters, employing different learning algorithms as base classifiers for noise detection. A noise detection ensemble E of size L is formed of a set of algorithms $\{A_1, \ldots, A_L\}$ that are used for noise identification. The individual classifiers can be combined to the final ensemble prediction using different combination rules [19]. Predictions of algorithms that return label outputs (like 'noise' and 'non-noise') can be combined using different voting schemes. Two most commonly used voting schemes for combining ensemble predictions are the following.

- *Majority* (*plurality*) *voting*: If more than half of the algorithms A_i from E identify an instance $\mathbf{x}$ as noisy, then the ensemble declares it as noisy.
- *Consensus* (*unitary*) *voting*: If all the algorithms A_i from E identify the instance $\mathbf{x}$ as noisy, then the ensemble declares it as noisy.

Let function δ be 1 for 'noisy' labels and 0 otherwise. Then the formal notation of the condition for noise identification of instance $\mathbf{x}$ by ensemble E using the majority voting scheme can be written as $\sum_{i=1}^{L} \delta(A_i(\mathbf{x})) > L/2$, and using the consensus voting scheme as $\sum_{i=1}^{L} \delta(A_i(\mathbf{x})) = L$.

2.2. Performance measures

Quantitative evaluation of noise detection methods requires to know which are the noisy instances in a dataset. In real-life

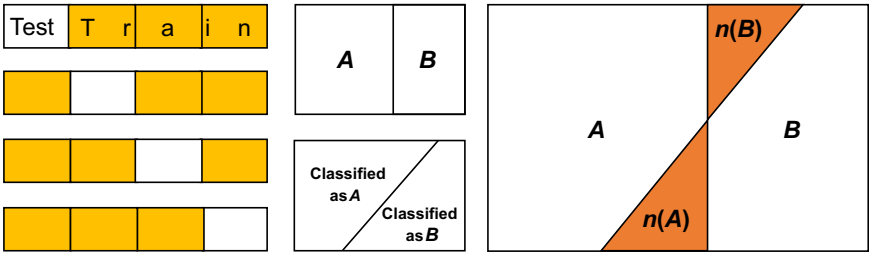

Fig. 1. Classification filtering using cross-validation. A and B are the class labels of instances in the test fold. The misclassified instances of A and B, denoted with $n(A)$ and $n(B)$, present the noise detected by the classification filter.

datasets this is achieved either by expert labeling or by random injection of errors into a dataset. A basic measure to evaluate the performance of noise detection is *precision*, defined as true positives divided by all the predicted positives (i.e., the percentage of correctly detected noisy instances among all the instances identified as noisy by the noise detection algorithm):

$$\text{Precision} = \frac{\text{number of true noisy instances detected}}{\text{number of all instances identified as noisy}}$$

Another useful measure is *recall*, which is defined as true positives divided by all the positives (i.e., the percentage of correctly detected noisy instances among all the noisy instances inserted into the dataset as random noise):

$$\text{Recall} = \frac{\text{number of true noisy instances detected}}{\text{number of all noisy instances in the dataset}}$$

To model a desired precision-recall tradeoff, the so-called *F*-measure combining precision and recall is used. The formula for computing the *F*-measure is the following:

$$F_\beta = (1+\beta^2)\frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}} \quad (1)$$

where for $\beta = 1$ we get the standard *F*-measure, also referred to as the F_1 score. By setting the β parameter, the user can assign more importance to either precision or recall in the calculation of the *F*-score.

2.3. Diversity measures

This section describes eleven diversity measures commonly used in the literature. They were designed to measure the difference in prediction errors made by the members of an ensemble. If ensemble members are diverse in terms of the errors they make, then their combination may reduce the total prediction error. The majority of diversity measures were proposed in the context of classification to indicate the diversity of an ensemble according to its classification performance. Since noise detection can be viewed as a classification problem of classifying instances as noisy and non-noisy (or regular), we adopted the same diversity measures as used for classification ensembles also for noise detection ensembles. Two sets of measures are introduced: pairwise and global (non-pairwise) diversity measures.

2.3.1. Pairwise diversity measures

Pairwise diversity measures are computed for each pair of classifiers or, in our case, noise detection algorithms from the set of L predictors that are used. The resulting $L(L-1)/2$ pairwise measures can be averaged to obtain an overall diversity value of the employed ensemble. Pairwise diversity measures for two noise detection algorithms A_i and A_j are calculated from the relative amounts of agreement and disagreement between the correct and incorrect predictions they make on a certain dataset. Table 1 shows the notation for the proportions of data instances in a dataset that were classified: correctly by both algorithms (a), correctly by A_i and incorrectly by A_j (b), incorrectly by A_i and correctly by A_j (c), and incorrectly by both algorithms (d).

Table 1
The relationship among the predictions of two noise detection algorithms. The values a, b, c and d represent relative amounts, hence it holds that $a+b+c+d=1$.

	A_j correct	A_j incorrect
A_i correct	a	b
A_i incorrect	c	d

Table 2
Pairwise diversity measures.

Diversity measure	Formula	Greater diversity
Correlation coefficient (*rho*)	$\rho_{i,j} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$	Smaller (↓)
Yule's Q statistic (Q)	$Q_{i,j} = \frac{ad-bc}{ad+bc}$	Smaller (↓)
Pairwise kappa (K_p)[a]	$\kappa_{i,j} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	Smaller (↓)
Disagreement (D)	$D_{i,j} = b+c$	Greater (↑)
Double Fault (D_F)	$DF_{i,j} = d$	Greater (↑)

[a] Correct formula taken from Fleiss [44], which differs from the one referenced in [19].

Table 3
Global diversity measures. Notation $\overline{p}$ stands for the average accuracy of all A_j, $a_j(\mathbf{x}_i)$ is 1 if A_j disagrees with the majority of predictors and 0 otherwise, N_c is the number of instances where the predictions of more than half of all A_j were correct, and $N_f = N - N_c$ is the number of instances where half or more of all A_j were incorrect (false).

Diversity measure	Formula	Greater div.
Entropy measure (E)	$E = \frac{1}{N}\frac{2}{L-1}\sum_{i=1}^{N} \min\{(Y_i), (L-Y_i)\}$	Greater (↑)
KW measure (*KW*)	$KW = \frac{1}{NL^2}\sum_{i=1}^{N} Y_i(L-Y_i)$	Greater (↑)
Interrater agreement (K_{np})	$\kappa = 1 - \frac{\frac{1}{L}\sum_{i=1}^{N} Y_i(L-Y_i)}{N(L-1)\overline{p}(1-\overline{p})}$	Smaller (↓)
Ambiguity (A)	$\overline{A} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} a_j(\mathbf{x}_i)$	Greater (↑)
'Good' diversity (D_g)	$D_g = \frac{1}{N_c L}\sum_{Y_{i,j} \geq \frac{L+1}{2}} (L - Y_{i,j})$	Greater (↑)
'Bad' diversity (D_b)	$D_b = \frac{1}{N_f L}\sum_{Y_{i,j} < \frac{L+1}{2}} Y_{i,j}$	Greater (↑)

In our work we use the pairwise diversity measures presented in [19], which are listed in Table 2. The last column in the table shows which (absolute) values of the measure indicate greater diversity.

Correlation coefficient (*rho*) is a measure of association between two binary predictors, also referred to as *Phi coefficient* [44]. *Yule's Q statistic* is a measure of association between the odds ratios of the algorithms, which indicate how much the odds of one algorithm making a correct (incorrect) prediction increase for cases where the other algorithm makes a correct (incorrect) prediction [45]. *Pairwise Kappa* (K_p) presents the ratio between (i) the excess of the observed agreement over the agreement by chance and (ii) the maximal excess over chance [44]. The two simplest pairwise diversity measures are D and D_F. The *Disagreement* measure (D) is the most intuitive diversity measure as it presents the proportion of data on which the two predictors disagree. The *Double Fault* measure (D_F) considers simultaneous errors to be more informative for diversity than the case when both predictors are correct [19].

2.3.2. Global diversity measures

Diversity measures that consider all predictors together and directly produce a single diversity value of the ensemble are referred to as global or non-pairwise measures. Let N be the number of instances in the observed dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=0}^{N}$, and let $y_{i,j} \in \{0, 1\}$ denote the correct or incorrect prediction of predictor A_j for $j \in \{1, \ldots, L\}$ on data instance $\mathbf{x}_i$ for $i \in \{1, \ldots, N\}$. The number of predictors that correctly recognize instance $\mathbf{x}_i$ are $Y_i = \sum_{j=1}^{L} y_{i,j}$. We investigate the global diversity measures presented in Table 3.

Entropy measure (E), as proposed in [22], assumes the ensemble to be most diverse when the number of correct (incorrect) member prediction is equal to $\lfloor L/2 \rfloor$ and the number of incorrect (correct) predictions is equal to $L - \lfloor L/2 \rfloor$, and least diverse when all predictions are either correct or incorrect. In [22] the authors presented the *KW*

measure as a modification of the *Kohavi–Wolpert variance* [46]. It measures the variability of predictions of a set of predictors $\{A_1, \ldots, A_L\}$ applied on a training set. The *Interrater agreement* measure (K_{np}) measures the level of agreement among different predictors while correcting for chance [19,44]. *Ambiguity* (A) measures the disagreement between individual predictors and the majority of the predictors over the given dataset [47]. "*Good*" *Diversity* (D_g) and "*Bad Diversity*" (D_b), proposed in [20], measure the disagreement between the prediction of A_j and the ensemble's prediction, for $j \in \{1, \ldots, L\}$: D_g measures the disagreement on the data instances where the ensemble is correct, and D_b on the instances where the ensemble is incorrect. Measures A, D_g and D_b assume a majority voting ensemble being used in their calculations.

3. Methodology

This section summarizes the aims of the paper, describes the experimental setting, and proposes an approach for empirical evaluation of the relation between ensemble diversity and noise detection performance.

3.1. Research hypothesis and goals

The effect of ensemble diversity on classification accuracy was extensively studied in the literature. In this paper the focus is different, we explore how diversity is related to the performance of ensembles for explicit noise detection. In this case, noise detection is not merely a preprocessing step in the learning phase of a potentially more accurate classification model, but is considered to be the main step in identifying noisy data instances for the purposes of data cleaning, improved data understanding, or semi-supervised outlier identification.

Our hypothesis is that ensemble diversity can be used as means for guiding the construction of well performing noise detection ensembles. More specifically, we want to show which diversity measures are more likely to indicate that certain ensembles can achieve higher noise recall or higher precision of noise detection. To this end, the goals of the paper are as follows.

- Develop a methodology for exploring the relation between ensemble diversity and noise detection performance.
  - (i) Set up a range of ensembles constructed from different noise detection algorithms and evaluate their performance on a set of datasets.
  - (ii) Select a list of ensemble diversity measures and compute them for each ensemble.
  - (iii) Empirically measure the relation between the computed ensemble diversity and the achieved noise detection performance.
- Examine whether ensemble diversity relates better to the performance of ensembles using the majority or the consensus voting scheme, which can be used in practice for semi-supervised or unsupervised noise detection, respectively.

3.2. Experimental setting

The experiments for empirically assessing the relation between ensemble diversity and noise detection performance were designed as follows. A total of 968 noise detection ensembles were used to detect different amounts of randomly injected class noise in ten standard UCI benchmark datasets. Their performance was evaluated in terms of recall, precision and the F-measure, separately for the majority and the consensus voting settings. The diversity of the ensembles was determined using the eleven diversity measures presented in Section 2.3. The diversity and performance results obtained for all the ensembles were used to calculate the correlation between the diversity and the performance measures, as will be described in the next section. The rest of this section provides details on the experimental setup in terms of ensemble construction, dataset selection and data preprocessing.

3.3. Ensembles

The noise detection ensembles employed in the experiments were constructed from different classification noise filters. This section presents the learning algorithms employed by the classification noise filters. A selection of ten learning algorithms was chosen from the algorithms available in the data mining environments Orange[1] [48] and Weka[2] [49], as presented in Table 4. In the selection we tried to include groups of learning algorithms that employ different approaches to learning. From the ten classification filters that were thereby obtained we constructed all possible combinations of 3–10 base predictors, i.e. ensembles of size 3–10, resulting in 968 different noise detection ensembles ($\sum_{m=3}^{10} \binom{10}{m} = 968$). The ensembles were used in two settings: using a majority and a consensus voting scheme.

3.4. Data

The diversity and noise detection performance of the constructed ensembles was measured on ten standard benchmark datasets from the UCI Machine Learning Repository [50]. In order to quantitatively evaluate the noise detection performance of the ensembles, it is required to know which instances in the datasets are noisy. Since this information was not available for the datasets, class noise was artificially introduced by switching labels of 5%, 10%, 15%, or 20% randomly selected instances of each dataset. For each dataset and each noise level the experiment was repeated ten times and the average diversity and performance values were reported.

In the experiments, the performance of noise detection ensembles is measured in terms of their ability to detect artificially injected noise. However, the original datasets may themselves contain noisy instances, which could be also detected as noisy by the ensembles. Given that in the evaluation we only count the correctly identified artificial noise, we have decided to make the datasets as "noiseless" as possible, by first eliminating inherent noise from the data. Thus, before introducing random noise, data cleaning was performed by a consensus of all ten classification filters, used to identify and eliminate the instances which are quite certainly noisy.

In data cleaning, the choice of the consensus filtering approach was preferred over majority filtering for two reasons. It is known that majority filtering may remove also many regular data instances, whereas consensus filtering is more conservative: (i) it removes only the instances for which we can be nearly certain that they represent inherent noise of the original datasets, and thereby (ii) considerably reduces the change of eliminating regular instances.

The selection of datasets is summarized in Table 5, where the original numbers of instances as well as the numbers of instances after the elimination of inherent noise in data cleaning are presented.

The cleaning of inherent noise resulted in the elimination of about 1–5% of instances in half of the datasets, and less than 0.5% of instances in the other datasets. It is interesting to note that even in two artificial datasets (*kr-vs-kp* and *tic-tac-toe*) few instances (2 and 3, respectively) were identified as noise by all ten classification noise filters. These instances are cases of tie game outcomes and are difficult to be distinguished as *win* or *no-win* outcomes (the two class labels used in the datasets). They are thus outliers or borderline examples of the concepts presented in their datasets,

[1] http://orange.biolab.si/
[2] http://www.cs.waikato.ac.nz/ml/weka/

Table 4
Learning algorithms from Orange [48] and Weka [49] used for classification noise filtering.

Orange	Weka
CN2 (rule learner)	J48 (decision tree learner)
kNN (nearest neighbor)	JRip
Naïve Bayes	Multilayer perceptron
Random forest	Random tree
SVM	SMO

Table 5
UCI datasets [50] used in the experiments.

Dataset	Instances	After cleaning	Attributes
breast-cancer-wisconsin	683	677 (99.1%)	9
credit	690	666 (96.5%)	15
diabetes	768	727 (94.7%)	8
ionosphere	351	348 (99.1%)	32
kr-vs-kp	3196	3194 (99.9%)	36
sick	3772	3767 (99.9%)	29
sonar	208	208 (100%)	60
spambase	4601	4563 (99.2%)	57
tic-tac-toe	958	955 (99.7%)	9
voting	435	429 (98.6%)	16

and were eliminated, despite of being regular instances of the datasets.

3.5. Relating diversity and performance

The goal of the study is to explore how the values of diversity measures change over different noise detection ensembles, and how these changes relate to the changes in noise detection performance of the respective ensembles. Therefore, for each ensemble E_j from the list $\{E_1,\ldots,E_K\}$ its diversity and performance are computed, and the corresponding series of all the diversity values is compared to the series of all the computed performance values. The correlation between these two series provides a measure of relation between ensemble diversity and noise detection performance. The formal definition of this experimental approach is described below.

1. For each diversity measure D_i, for $i \in \{1,\ldots,M\}$, we calculate a series of values $d_{i,j}$ for $j \in \{1,\ldots,K\}$, for the list of K evaluated ensembles E_j (for illustration, the series of diversity values are shown in Fig. 10 of the Appendix).
2. For noise detection performance measures precision, recall and the F-measure, we calculate separate series $\{p_j\}_{j=1}^{K}$, $\{r_j\}_{j=1}^{K}$ and $\{f_j\}_{j=1}^{K}$, for the corresponding K ensembles (illustrated in Figs. 11 and 12 in the Appendix, for the majority and consensus voting scheme, respectively).
3. The relation of changes in ensemble diversity to the changes in performance can now be modeled by the correlation among the M series $\mathcal{S}_D = \{\{d_{i,j}\}_{j=1}^{K}\}_{i=1}^{M}$ and the three series $\mathcal{S}_P = \{\{p_j\}_{j=1}^{K}, \{r_j\}_{j=1}^{K}, \{f_j\}_{j=1}^{K}\}$.
4. For each pair of series $(\{d_{i,j}\}_{j=1}^{K}, \{s_j\}_{j=1}^{K})$ from the $S_D \times S_P$, the Spearman's rank correlation coefficient [51] is calculated as

$$\rho_{i,s} = \frac{\sum_{j=1}^{K}(d'_{i,j}-\overline{d'_i})(s'_j-\overline{s}')}{\sqrt{\sum_{j=1}^{K}(d'_{i,j}-\overline{d'_i})^2 \sum_{j=1}^{K}(s'_j-\overline{s}')^2}},$$

where $d'_{i,j}$ and s'_j denote the ranks of values $d_{i,j}$ and s_j in their respective series, and the notations $\overline{d'_i}$ and $\overline{s}'$ stand for the average (arithmetic mean) ranks in the series.

The Spearman's rank correlations coefficient is basically the Pearson's correlation coefficient [52] calculated on the ranks of the values in the series. We chose to use this transformation as it enables us to measure the more general monotone relationship between the two series, rather than measuring only the linear relationship.

4. Results

To evaluate ensemble diversity with regard to their class noise detection performance, we obtained eleven series of values for the given diversity measures (introduced in Section 2.2), and six series of values for the performance measures: three (precision, recall and F) for each of the two voting schemes (majority and consensus). These series show the changes of diversity and noise detection performance over the set of 968 evaluated ensembles.

To assess the relation between diversity and performance of noise detection ensembles, we calculated the Spearman's correlation coefficients (as described in Section 3.5) for all pairs of one diversity measure and one performance measure, separately for each of the two voting schemes. The average correlations[3] between the values of diversity measures and performance measures, calculated on ten datasets with four different levels of injected class noise, are presented in Fig. 2. The correlation results over all datasets for separate noise levels (5%, 10%, 15% and 20%) are presented in the Appendix in Figs. 13, 14, 15 and 16, respectively.

The figures present the magnitude and the orientation (sign) of the correlations between diversity measures and performance measures. The results, averaged over all noise levels and shown in Fig. 2, can now be interpreted for each of the two voting schemes.

- Noise detection results obtained by majority voting (shown on the left hand side in Fig. 2) are on average weakly negatively correlated with the results of diversity measures. For most relations the results show (very) weak negative correlations lower than 0.4. Moderate negative correlation is observed for the D_F measure with all the performance measures, for D, E and D_g with noise recall, and for D_b with the precision of noise detection. Weak positive correlation is observed only between D_b and noise recall. The relatively high standard deviations show that the correlations are also very domain dependent, but not noise-level dependent, as they are very similar for all the noise levels. Generally the results for ensemble-based noise detection using majority voting imply, although weakly, that the noise detection ensembles used in our experiments achieve higher noise detection performance when they are less diverse, i.e., when member predictions are more similar.
- In the case of the consensus voting scheme for class noise identification (shown on the right hand side of Fig. 2), the correlations between the diversity measures and noise detection performance are significantly higher. Over all diversity measures, except D_F and E, the same pattern can be observed that a diversity measure is inversely correlated with recall and precision, e.g., D_b is negatively correlated with recall and positively correlated with precision. The same holds for precision and the F-score. Very strong negative correlations with noise detection performance can be observed for KW, A and D_g with recall and with the F-score. On the other hand, moderately

[3] Note that for four diversity measures—Correlation (*rho*), Q statistic (Q), Pairwise kappa (K_p), and Interrater agreement (K_{np})—the calculated correlation coefficients were multiplied by -1 in order to avoid potential misinterpretation of the obtained results. The reason is that while typically higher values of diversity measures mean higher diversity, the opposite is true for these four measures, where lower values indicate higher diversity.

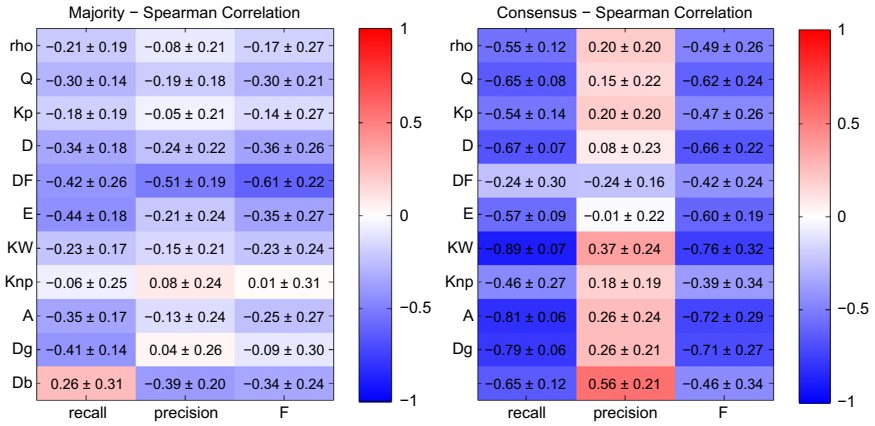

Fig. 2. Average correlations between diversity measures and noise detection performance over ten experimental datasets at all four levels of injected class noise. Results presented for ensembles used in the majority (left) and consensus (right) voting scheme. Note that, for sample size $n=968$ and significance level $\alpha=0.05$, correlation values $|\rho| > 0.06$ are statistically significant according to the Student's t-test.

positive correlation for D_b and KW with precision can be observed. It is interesting to note that diversity measures A, D_G and D_b, which achieve high (absolute) correlations with consensus-based noise detection, assume the use of a majority voting scheme in the calculation of the diversity value. Additionally, it is encouraging to notice that the highest correlations are accompanied by relatively lower standard deviations, except for the F-measure where the high standard deviations are due to significantly lower correlation results at noise level 5% (see Fig. 13 in the Appendix).

For the majority voting scheme, the above results may seem counter-intuitive, while for the consensus voting scheme the results can easily be explained. As for the latter, reaching consensus of diverse classifiers is hard, but when a consensus of a very heterogeneous classifier is reached, one can trust that the detected noisy instances indeed represent noise in the data (which is reflected in high precision of noise detection). This explains the high correlation between diversity and performance (precision) for the consensus voting scheme. On the other hand, the diversity property is not so important in the much less restrictive majority voting scheme. Consider an illustrative example for the majority voting case. Suppose that an ensemble consists of 7 classifiers—4 diverse and 3 very similar—and that an instance is identified as noisy by 4 votes out of 7 votes. Given that the ensemble includes three similar classifiers, its diversity will not be evaluated as high, even though the correctly identified noisy instance might have been identified by the four diverse classifiers. As shown in this example, the overall diversity of the ensemble may not reflect the nature of the subset of classifiers which have identified the instance as noisy in the majority voting scheme, whereas in the consensus voting scheme the ensemble diversity measure indeed reflects the nature of the set of all classifiers that have identified the instance as noisy.

In Fig. 2, the absolute correlations between ensemble diversity and noise detection performance range from weak and moderate (0.3–0.5) for the majority voting setting, to strong and very strong (0.7–0.9) for the consensus voting setting. Therefore we wanted to test whether there are any statistically significant differences among the correlations of the diversity measures with a certain performance measure. To this end, we used the Friedman test ($\alpha=0.05$) with the Nemenyi post hoc test, which can be used for comparing the performance of several algorithms over multiple datasets, as suggested in [53]. The test compares the average ranks of the algorithms (i.e., diversity measures achieving a correlation with a certain performance measure) over all experimental datasets. The results of this statistical test can be visualized by the *critical difference* (CD) diagrams, also suggested in [53].

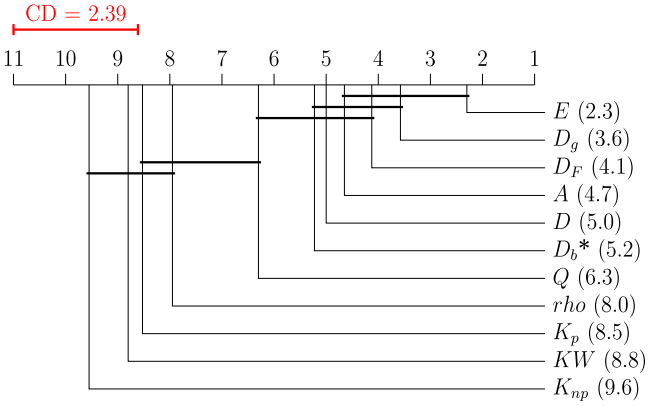

Fig. 3. CD diagram for the correlations of diversity measures with recall in the majority voting setting. Note that unlike other measures D_b is positively correlated with recall.

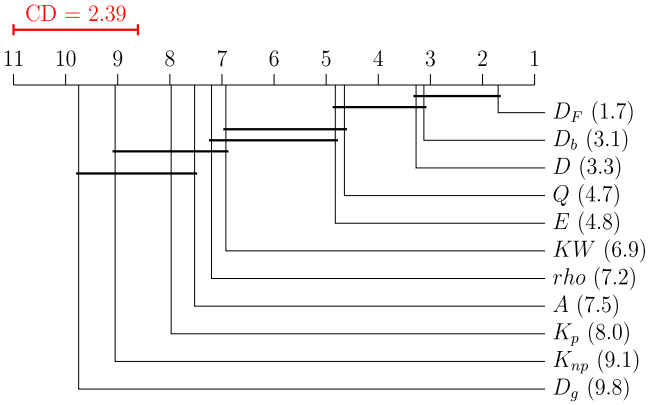

Fig. 4. CD diagram for the correlations of diversity measures with precision in the majority voting setting.

Statistical differences between the correlations of diversity measures with a certain performance measure are presented in Figs. 3–5 for the majority voting setting, and in Figs. 6–8 for the consensus voting setting. The CD diagrams show the average ranks of the absolute correlation results for the different diversity measures with a given performance measure achieved over all experimental datasets. The critical difference (CD) indicates the difference in ranks that has to

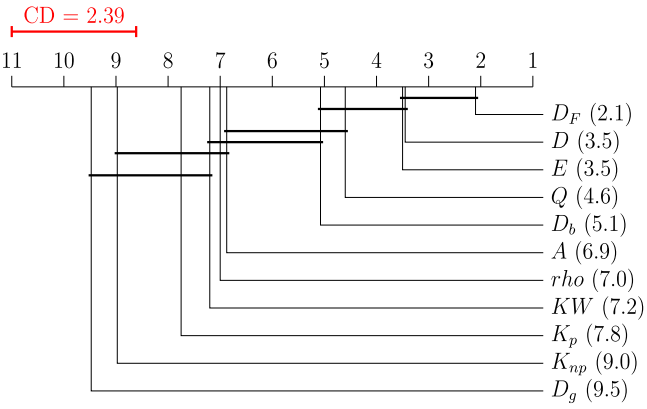

Fig. 5. CD diagram for the correlations of diversity measures with the *F*-measure in the majority voting setting.

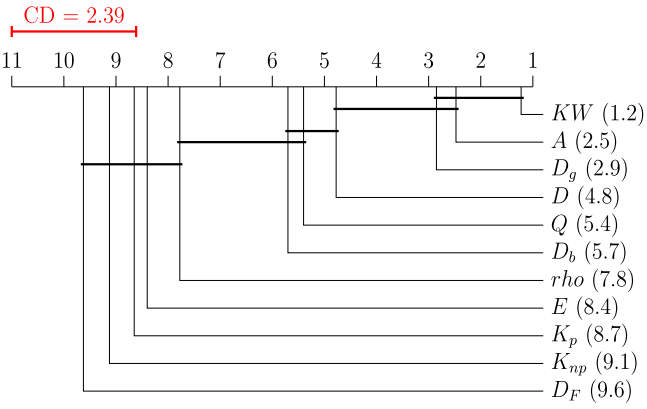

Fig. 6. CD diagram for the correlations of diversity measures with recall in the consensus voting setting.

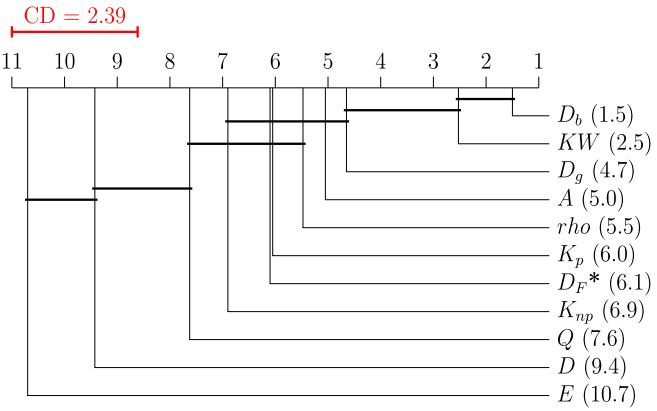

Fig. 7. CD diagram for the correlations of diversity measures with precision in the consensus voting setting. Note that unlike other measures D_F is negatively correlated with precision.

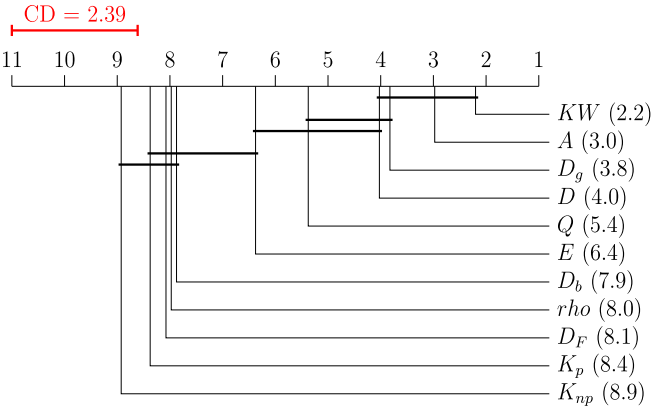

Fig. 8. CD diagram for the correlations of diversity measures with the *F*-measure in the consensus voting setting.

be exceeded in order to be statistically significantly different. If the difference of average ranks of two correlation results is smaller than the value of CD, then they are—according to this test—not statistically significantly different (such an outcome is shown by a thick black line connection in the diagram).

The results of statistical significance have to be interpreted together with the results from Fig. 2 showing the magnitude of the correlations. For example, consider the CD diagram for the majority voting setting for the correlations of diversity measures with recall presented in Fig. 3. The *E* measure is on average the one which most often achieves the highest (absolute) correlation with recall (average rank 2.3). However, it is not statistically significantly different from the correlations with recall observed for D_g, D_F and *A*. But it is statistically significantly different from the correlations for the other seven measures. In this sense the CD diagrams enable us to see the statistical significance of the observed correlations, as well as the average ranks of correlations achieved for a pair of a diversity measure and a performance measure in the experimental evaluation. The CD diagrams for all the performance measures in both voting settings show the same order of most highly correlated pairs as in Fig. 2, but provide an additional perspective when selecting a diversity measure that would be most likely to indicate which ensemble would achieve good noise detection performance.

The correlation analysis experiments provide promising results regarding the relation between the diversity of ensembles and their noise detection performance. However, the question remains whether the relation is sufficiently indicative to be used for selecting best performing noise detection ensembles. In our attempt to answer this question we have examined the overlap between the best performing noise detection ensembles and the most diverse or most non-diverse ensembles as suggested by the prevailing sign of correlations between diversity measures and the observed performance measure. This was achieved by measuring the overlap with the so-called *precision at k* method [54], a commonly used metric in recommender systems, which is in our case calculated as follows.

1. Make a list l_P of all ensembles sorted according to their performance measure *P* from best to worst.
2. Make another list l_D of all ensembles sorted either:
   (a) from their highest to their lowest diversity values if most diversity measures are positively correlated with performance measure *P*, or
   (b) from their lowest to their highest diversity values if most diversity measures are negatively correlated with performance measure *P*.
3. Compare the top *k* elements of the lists l_P and l_D for $k \in \{1, \ldots, 968\}$ and return the relative size of the overlap $r(k) = |l_P(k) \bigcap l_D(k)| / k$.

This was calculated for all pairs of one performance measure and one diversity measure. Note that in our case the lists l_D were ordered according to 2(b) for both voting settings and all performance measures, except for precision in the consensus voting setting where positive correlation between diversity and performance was observed. The changes in relative overlaps for recall, precision and the *F*-measure in the case of majority-based and consensus-based noise detection on data with 10% class noise are presented in Fig. 9.

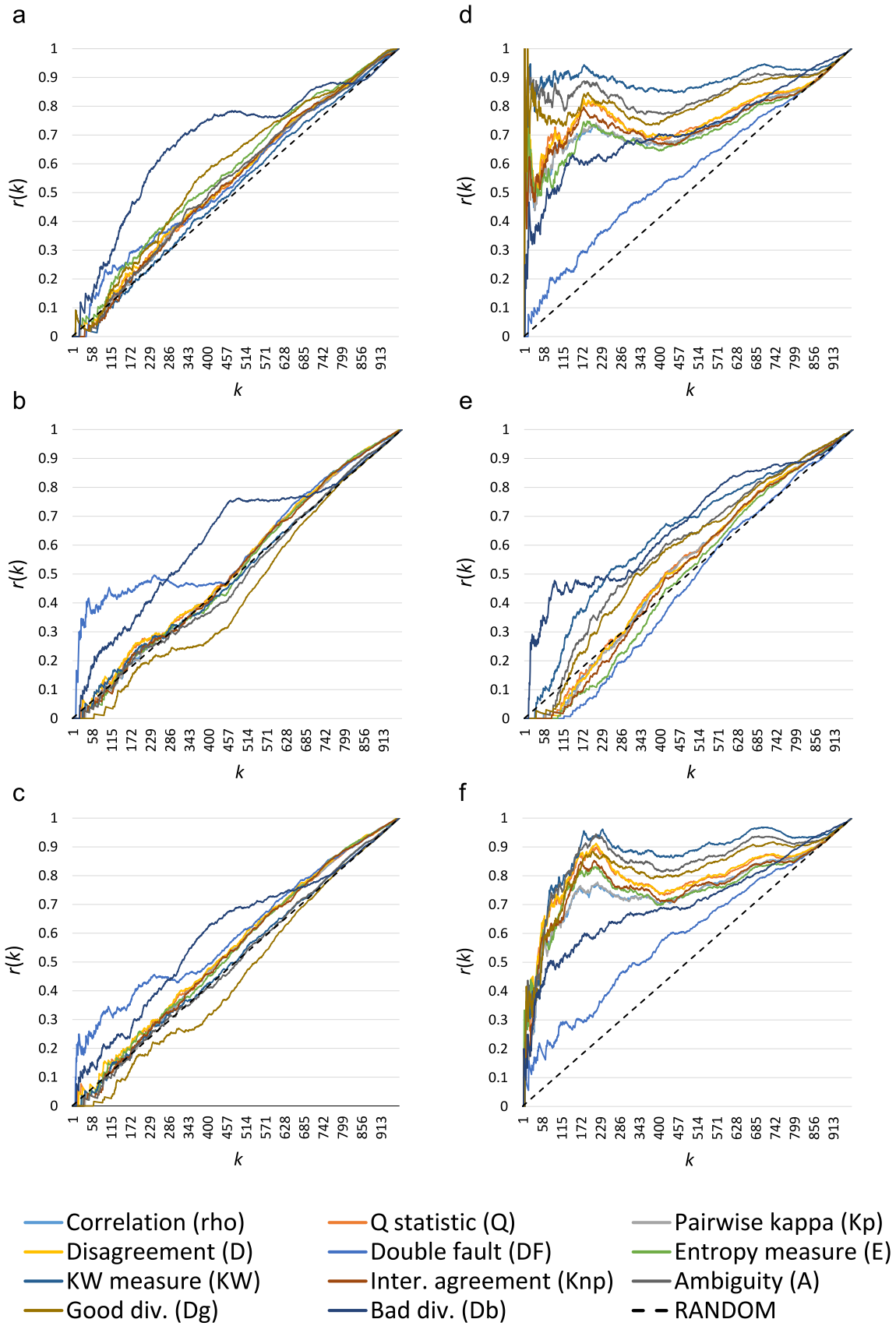

Fig. 9. Relative overlaps $r(k)$ between top k ensembles in terms of performance values and diversity values for the majority and consensus voting setting.

The dashed diagonal in the charts of Fig. 9 indicates the average expected overlap between two unsorted or randomly ordered lists. In terms of correlation the dashed diagonal can be interpreted as a curve showing no correlation between the noise detection performance and the diversity of an ensemble. According to the decreasing or increasing ordering of l_D, high positive or high

negative correlation between performance and diversity results in curves above the diagonal, whereas curves below the diagonal indicate the reverse correlation according to the way in which the lists l_D were ordered. Curves that are close to the diagonal for all values of k indicate low or no correlation.

Compared to the presentation of correlation coefficients in Fig. 2, the advantage of overlap visualization is that it enables to identify interval(s) where the performance and diversity measures are more related. For example, in Fig. 9(f) the highest *F*-scores are tightly related to the lowest diversity values of the *KW* measure, Ambiguity *A* and the Disagreement measure *D*, as these top performance and lowest diversity results were obtained by almost the same ensembles (90%). However for $k > 200$ the relation between the performance and these diversity measures decreases. In other words, among the ensembles achieving the highest *F*-scores is also a great majority of ensembles with the lowest diversity values, however for lower performance and higher diversity values this relation becomes weaker.

The results in charts (a)–(c) of Fig. 9 show that although some diversity measures are weakly negatively correlated with the noise detection performance of majority voting ensembles, the information about ensembles' diversity cannot be effectively used for identifying the best performing noise detection ensembles. Nevertheless, best choices of an indicative diversity measure may be the following. To achieve high recall, high D_b values may be most informative from a certain point on (chart (a) for $k > 100$); to achieve high precision, following the lowest D_F values would be likely to result in the discovery of 40–50% of most precise noise detection ensembles (chart (b) for $50 < k < 250$); and to achieve high *F*-scores, lowest D_F values are even less likely to capture a reasonable amount of best *F*-score achieving ensembles (chart (c)).

On the other hand, when using consensus voting ensembles, as presented in charts (d)–(f) of Fig. 9, several diversity measures may be used to select a large part of noise detection ensembles achieving high noise recall and high *F*-scores. The ensembles achieving best recall can be identified by following the lowest diversity values of measures *KW*, *A* and D_g, namely 80–90% overlap among the 50 top-ranked ensembles (chart (d)). Also for the *F*-measure, following the ordering of ensembles with lowest values of *KW*, *A* and *D* results in 80–95% overlap for $100 < k < 200$ (chart (e)). In terms of precision, the D_b measure offers the most promising (almost 50%) chance of encountering most precise noise detection ensembles if following the lead of highest D_b values (chart (f)).

5. Conclusions

This paper presents a study in ensemble-based class noise detection. It investigates the relation between the diversity of heterogeneous ensembles of noise detection algorithms and their class noise detection performance, with the hypothesis that ensemble diversity may be used as guidance for selection of well performing noise detection ensembles.

The relation between ensemble diversity and noise detection performance was assessed empirically through the correlations among the series of diversity values and performance values obtained for all the evaluated ensembles on a number of UCI datasets with different levels of randomly injected class noise. The hypothesis that ensemble diversity may be used as guidance for selection of well performing noise detection ensembles was experimentally tested in two different settings of ensemble-based noise detection, using the majority voting scheme and the consensus voting scheme. The results are summarized for both settings, showing that the hypothesis was confirmed only when using the consensus voting scheme.

We first summarize the negative results for the majority voting setting. Noise detection ensembles using the majority voting scheme are known to achieve high recall of noisy instances, but typically falsely identify a lot of regular instances as noisy. Therefore the majority voting scheme is in practice well suited for expert-guided noise detection where high recall of irregular data instances is more important than high noise detection precision. The experimental results show that in the case of ensemble-based class noise detection using majority voting, ensemble diversity does not positively correlate with noise detection performance: all the diversity measures agree that less diverse ensembles lead to better noise detection performance.[4] Additional experiments showed that even the highest correlations between diversity values and noise detection performance in the majority voting scheme provide only little guidance in selecting well performing noise detection ensembles.

In contrast to majority voting ensembles, noise detection ensembles that use a consensus voting scheme tend to be very precise in finding noise, but typically their noise recall is not as good, leaving more noisy instances undetected. Consensus-based ensembles are suitable for unsupervised noise detection in use cases where only the most significant noise should be removed and where the detection of false noise should be avoided. The experiment with consensus voting ensembles showed significantly higher correlations between ensemble diversity and noise detection performance. The results show that more diverse ensembles tend to achieve higher precision of class noise detection, whereas less diverse ensembles tend to achieve higher recall of noise detection and higher *F*-scores. Taking this into account, we have shown that selected diversity measures can be used as guidance for choosing well performing noise detection ensembles using the consensus voting scheme, given that the *KW* measure and Ambiguity *A* are strongly correlated with noise recall and the *F*-measure, and the 'Bad' diversity measure with the precision of noise detection.

In future work, we plan to implement all the analytical methods presented in this paper into an open source data mining platform. This will enable public accessibility and reuse of the presented analytic process for assessing the diversity and the performance of noise detection ensembles, as well as the relation between them. Moreover, we plan to explore the relation of ensemble diversity and noise detection performance in voting schemes that span the gap between majority and consensus voting. This idea is motivated by the success of homogeneous noise filtering ensembles using a new 'high-agreement' voting scheme implemented in the recently developed High-Agreement Random Forest noise filtering ensemble [32]. By extending the range of voting schemes and evaluation domains we hope to further contribute to the art of selecting best performing noise detection ensembles.

Acknowledgments

This work was supported in part by the European Commission under the FP7 project MULTIPLEX (Foundational Research on MULTIlevel comPLEX networks and systems, Grant no. 317532), and by the Slovenian Research Agency programme Knowledge Technologies (Grant no. P2-103).

Appendix

This section includes two types of addition figures. First, are the figures illustrating the series of diversity and performance values used in the computation of correlation between ensemble

[4] Except in the case of the 'Bad' diversity D_b measure, where a weak tendency is observed that according to this measure more diverse ensembles achieve higher recall.

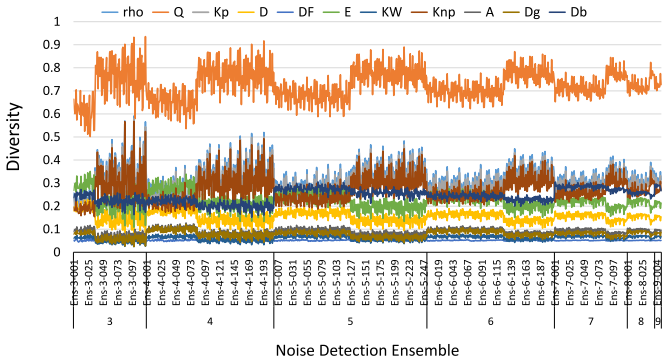

Fig. 10. Averages of 11 diversity measures on 968 ensembles over ten datasets.

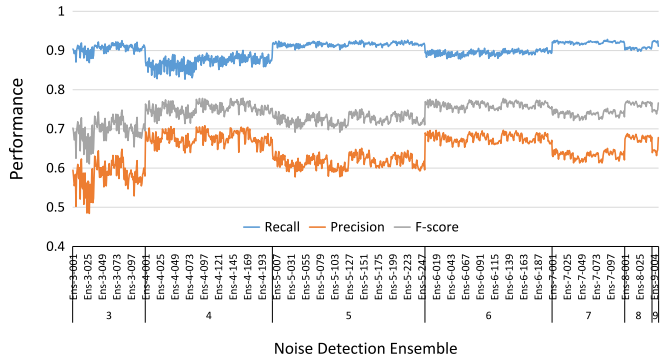

Fig. 11. Average performance of 968 ensembles with majority voting over ten datasets.

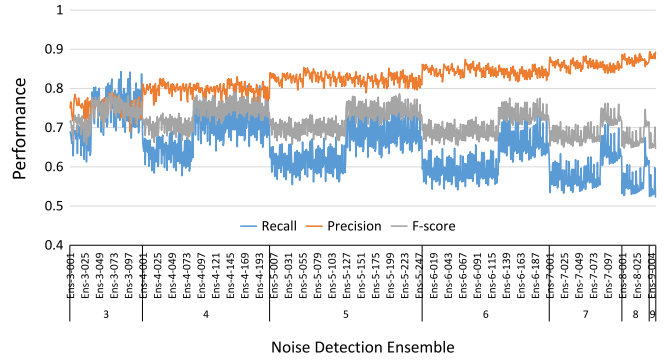

Fig. 12. Average performance of 968 ensembles with consensus voting over ten datasets.

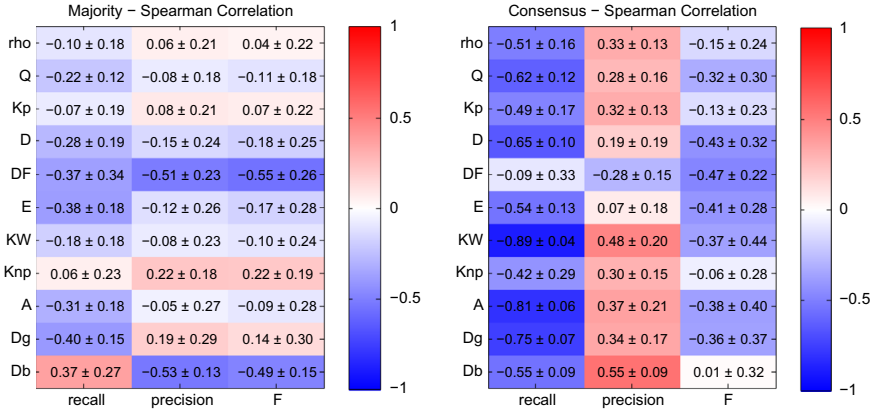

Fig. 13. Average correlations of diversity vs. performance on data with 5% class noise.

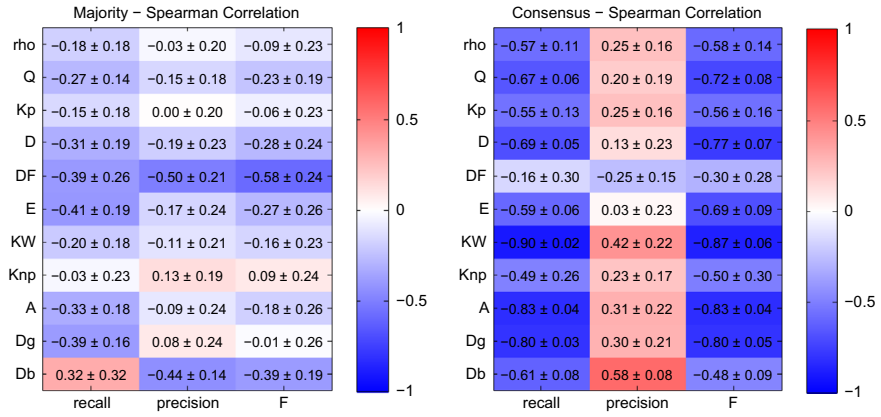

Fig. 14. Average correlations of diversity vs. performance on data with 10% class noise.

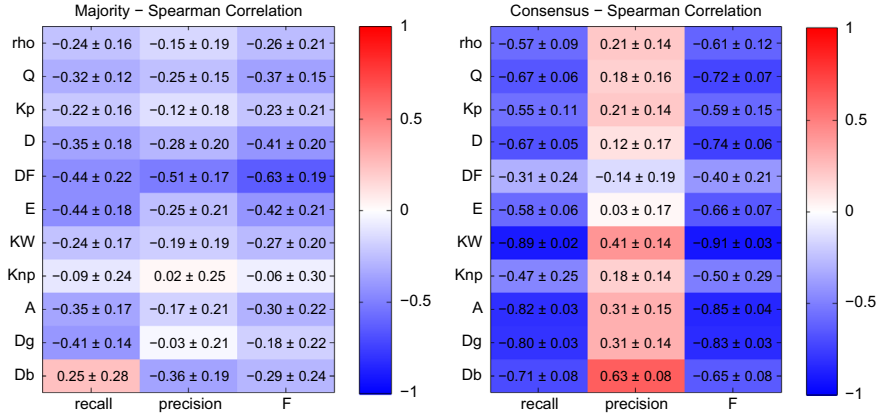

Fig. 15. Average correlations of diversity vs. performance on data with 15% class noise.

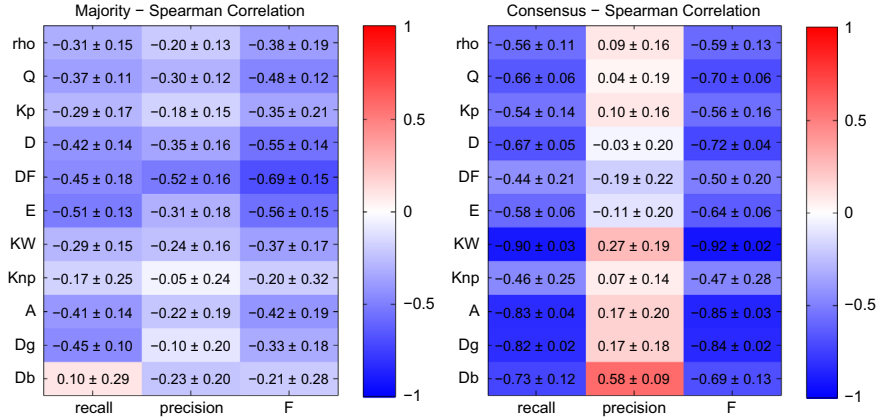

Fig. 16. Average correlations of diversity vs. performance on data with 20% class noise.

diversity measures and noise detection performance. Second, are the Figs. 13, 14, 15 and 16 of correlation results between diversity and performance measures on all datasets at separate noise levels 5%, 10%, 15% and 20%, respectively.

Figs. 10–12 present the average diversities and performances of the 968 ensembles using the majority and consensus schemes, achieved on ten datasets with 10% class noise. This specific noise level was selected just to show the movement of the series of diversity values and series of performance values that will be used to measure the relation among ensemble diversity and noise detection performance. As the behavior of the series at other noise levels is similar, their presentation was skipped. In the figures the noise detection ensembles are grouped by their size and ordered according to the joint of their members' names. For ease of presentation were the ensemble names in the figures encoded as "Ens-'ensemble size'-'sequential number'".

References

[1] X. Zhu, X. Wu, Class noise vs. attribute noise: a quantitative study of their impacts, Artif. Intell. Rev. 22 (2004) 177–210.
[2] B. Frénay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE Trans. Neural Netw. Learn. Syst. (2013) 1–25.
[3] T.G. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems, 2000, pp. 1–15.
[4] R. Polikar, Ensemble learning, Scholarpedia 4 (1) (2009) 2776.
[5] G. Brown, J.L. Wyatt, R. Harris, X. Yao, Diversity creation methods: a survey and categorisation, Inf. Fus. 6 (1) (2005) 5–20.
[6] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, Limits on the majority vote accuracy in classifier fusion, Pattern Anal. Appl. 6 (1) (2003) 22–31.

[7] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (1990) 197–227.
[8] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[9] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[10] K.-W. Hsu, J. Srivastava, Diversity in combinations of heterogeneous classifiers, In: T. Theeramunkong, B. Kijsirikul, N. Cercone, T.B. Ho, (Eds.), Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference, PAKDD, Lecture Notes in Computer Science, vol. 5476, 2009, Springer, Berlin, Heidelberg, 923–932.
[11] M. Gashler, C.G. Giraud-Carrier, T.R. Martinez, Decision tree ensemble, small heterogeneous is better than large homogeneous, In: M.A. Wani, X. wen Chen, D. Casasent, L.A. Kurgan, T. Hu, K. Hafeez, (Eds.), Proceedings of the Seventh International Conference on Machine Learning and Applications, ICMLA, 2008, IEEE Computer Society, Washington, 900–905.
[12] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, Ensemble selection from libraries of models, In: C.E. Brodley, (Ed), Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), ACM International Conference Proceeding Series, vol. 69, 2004, ACM, New York, 18–26.
[13] R. Caruana, A. Munson, A. Niculescu-Mizil, Getting the most out of ensemble selection, in: Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), 2006, IEEE Computer Society, pp. 828–833.
[14] I. Partalas, G. Tsoumakas, I.P. Vlahavas, Focused ensemble selection: a diversity-based method for greedy ensemble selection, in: M. Ghallab, C.D. Spyropoulos, N. Fakotakis, N.M. Avouris (Eds.), ECAI, Frontiers in Artificial Intelligence and Applications, vol. 178, IOS Press, Amsterdam, 2008, pp. 117–121.
[15] I. Partalas, G. Tsoumakas, I.P. Vlahavas, Pruning an ensemble of classifiers via reinforcement learning, Neurocomputing 72 (7–9) (2009) 1900–1909.
[16] N. Li, Y. Yu, Z.-H. Zhou, Diversity regularized ensemble pruning, in: P.A. Flach, T.D. Bie, N. Cristianini (Eds.), ECML/PKDD (1), Lecture Notes in Computer Science, vol. 7523, Springer, Berlin, Heidelberg, 2012, pp. 330–345.
[17] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.
[18] C.J. Merz, Using correspondence analysis to combine classifiers, Mach. Learn. 36 (1–2) (1999) 33–58.
[19] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, New Jersey, USA, 2004.
[20] G. Brown, L.I. Kuncheva, "Good" and "bad" diversity in majority vote ensembles, in: N.E. Gayar, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, 9th International Workshop, (MCS 2010), Lecture Notes in Computer Science, vol. 5997, Springer, Berlin, Heidelberg, 2010, pp. 124–133.
[21] I. Khamassi, M. Sayed-Mouchaweh, M. Hammami, K. Ghdira, Ensemble classifiers for drift detection and monitoring in dynamical environments, in: Proceedings of the Annual Conference of the Prognostics and Health Management Society, 2013, pp. 199–212.
[22] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51 (2) (2003) 181–207.
[23] S. Wang, X. Yao, Diversity analysis on imbalanced data sets by using ensemble models, in: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM '09, 2009, pp. 324–331.
[24] S. Whalen, G. Pandey, A comparative analysis of ensemble classifiers: case studies in genomics, In: H. Xiong, G. Karypis, B.M. Thuraisingham, D.J. Cook, X. Wu, (Eds.), Proceedings of the 13th IEEE International Conference on Data Mining (ICDM), 2013, IEEE, New York, 807–816.
[25] A.F. Neto, A.M.P. Canuto, T.B. Ludermir, Using good and bad diversity measures in the design of ensemble systems: a genetic algorithm approach, in: IEEE Congress on Evolutionary Computation, IEEE, New York, 2013, pp. 789–796.
[26] G. Fumera, F. Roli, A theoretical and experimental analysis of linear combiners for multiple classifier systems, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 942–956.
[27] E.K. Tang, P.N. Suganthan, X. Yao, An analysis of diversity measures, Mach. Learn. 65 (1) (2006) 247–271.
[28] D. Gamberger, N. Lavrač, C. Grošelj, Experiments with noise filtering in a medical domain, in: Proceedings of the 16th International Conference on Machine Learning – ICML, Morgan Kaufmann, San Francisco, 1999, pp. 143–151.
[29] D. Gamberger, N. Lavrač, G. Krstačić, Active subgroup mining: a case study in a coronary heart disease risk group detection, Artif. Intell. Med. 28 (2003) 27–57.
[30] A. Loureiro, L. Torgo, C. Soares, Outlier detection using clustering methods: a data cleaning application, in: Proceedings of the Data Mining for Business Workshop, 2004.
[31] J.D. Van Hulse, T.M. Khoshgoftaar, H. Huang, The pairwise attribute noise detection algorithm, Knowl. Inf. Syst. 11 (2) (2007) 171–190.
[32] B. Sluban, D. Gamberger, N. Lavrač, Ensemble-based noise detection: noise ranking and visual performance evaluation, Data Min. Knowl. Discov. 28 (2) (2014) 265–303.
[33] C.E. Brodley, M.A. Friedl, Identifying mislabeled training data, J. Artif. Intell. Res. 11 (1999) 131–167.
[34] S. Verbaeten, A. Van Assche, Ensemble methods for noise elimination in classification problems, in: T. Windeatt, F. Roli (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 2709, Springer, Berlin, Heidelberg, 2003, pp. 317–325.
[35] T.M. Khoshgoftaar, S. Zhong, V. Joshi, Enhancing software quality estimation using ensemble-classifier based noise filtering, Intell. Data Anal. 9 (1) (2005) 3–27.
[36] T.M. Khoshgoftaar, V.H. Joshi, N. Seliya, Detecting noisy instances with the ensemble filter: a study in software quality estimation, Int. J. Softw. Eng. Knowl. Eng. 16 (1) (2006) 53–76.
[37] S. Zhong, W. Tang, T. M. Khoshgoftaar, Boosted Noise Filters for Identifying Mislabeled Data, Technical Report, Department of Computer Science and Engineering, Florida Atlantic University, 2005.
[38] A. Karmaker, S. Kwek, A boosting approach to remove class label noise, Int. J. Hybrid Intell. Syst. 3 (3) (2006) 169–177.
[39] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation forest, in: Proceedings of the 8th IEEE International Conference on Data Mining, IEEE Computer Society, Washington, 2008, pp. 413–422.
[40] B. Sluban, D. Gamberger, N. Lavrač, Advances in class noise detection, in: H. Coelho, R. Studer, M. Wooldridge (Eds.), Proceedings of the 19th European Conference on Artificial Intelligence, ECAI 2010, Frontiers in Artificial Intelligence and Applications, vol. 215, IOS Press, Amsterdam, Netherlands, 2010, pp. 1105–1106.
[41] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation-based anomaly detection, Trans. Knowl. Discov. Data 6 (1) (2012) 3.
[42] K. Noto, C.E. Brodley, D.K. Slonim, Anomaly detection using an ensemble of feature models, In: G.I. Webb, B. Liu, C. Zhang, D. Gunopulos, X. Wu, (Eds.), Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), 2010, IEEE Computer Society, Washington, 953–958.
[43] K. Noto, C.E. Brodley, D.K. Slonim, Rac: a feature-modeling approach for semi-supervised and unsupervised anomaly detection, Data Min. Knowl. Discov. 25 (1) (2012) 109–133.
[44] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions, 3rd edition, Wiley-Interscience, New Jersey, USA, 2003.
[45] G. Yule, On the association of attributes in statistics, Philos. Trans. R. Soc. Lond. 194 (1900) 257–319.
[46] R. Kohavi, D.H. Wolpert, Bias plus variance decomposition for zero-one loss functions, in: Proceedings of the 13th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, 1996, pp. 275–283.
[47] G. Zenobi, P. Cunningham, Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error, in: L. D. Raedt, P.A. Flach (Eds.), ECML, Lecture Notes in Computer Science, vol. 2167, Springer, Berlin, Heidelberg, 2001, pp. 576–587.
[48] J. Demšar, T. Curk, A. Erjavec, Črt GorupT. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, B. Zupan, Orange: data mining toolbox in python, J. Mach. Learn. Res. 14 (2013) 2349–2353.
[49] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newslet. 11 (1) (2009) 10–18.
[50] K. Bache, M. Lichman, UCI Machine Learning Repository. ⟨http://archive.ics.uci.edu/ml⟩, 2013.
[51] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101.
[52] K. Pearson, Note on regression and inheritance in the case of two parents, Proc. R. Soc. Lond. 58 (1895) 240–242.
[53] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[54] V. Raghavan, P. Bollmann, G.S. Jung, A critical investigation of recall and precision as measures of retrieval system performance, ACM Trans. Inf. Syst. 7 (3) (1989) 205–229.

Borut Sluban received his BSc in Applied mathematics from the University of Ljubljana, Slovenia, and his PhD in Information and Communication Technologies from Jožef Stefan International Postgraduate School, Ljubljana, Slovenia. He is employed as a research assistant at the Department of Knowledge Technologies at the Jožef Stefan Institute, Ljubljana, Slovenia. His research is in the field of ensemble based noise and outlier detection, performance evaluation, and content extraction from large scale textual data streams. His current research interests include data/text mining, information retrieval, content extraction, and knowledge discovery from data streams and induced networks.

Nada Lavrač is the Head of Department of Knowledge Technologies at Jožef Stefan Institute, Ljubljana, Slovenia. She is also the Professor at the Jožef Stefan International Postgraduate School in Ljubljana and at the University of Nova Gorica. Her main research interests are in Knowledge Technologies, with special interests in machine learning, data mining, text mining, and computational creativity. Her special interest is in supervised descriptive rule induction and particularly subgroup discovery, where the research goal is to automatically induce descriptive rules from class labeled data, stored either in simple tabular format or in complex relational databases. Areas of her applied research include data mining applications in medicine, health care and bioinformatics. She is the author of several books, including the recently published book Foundations of Rule Learning, Springer, 2012.