Diversity in Combinations of Heterogeneous Classifiers

Kuo-Wei Hsu and Jaideep Srivastava

University of Minnesota, Minneapolis, MN, USA
{kuowei, srivastava}@cs.umn.edu

Abstract. In this paper, we introduce the use of combinations of heterogeneous classifiers to achieve better diversity. Conducting theoretical and empirical analyses of the diversity of combinations of heterogeneous classifiers, we study the relationship between heterogeneity and diversity. On the one hand, the theoretical analysis serves as a foundation for employing heterogeneous classifiers in Multi-Classifier Systems or ensembles. On the other hand, experimental results provide empirical evidence. We consider synthetic as well as real data sets, utilize classification algorithms that are essentially different, and employ various popular diversity measures for evaluation. Two interesting observations will contribute to the future design of Multi-Classifier Systems and ensemble techniques. First, the diversity among heterogeneous classifiers is higher than that among homogeneous ones, and hence using heterogeneous classifiers to construct classifier combinations would increase the diversity. Second, the heterogeneity primarily results from different classification algorithms rather than the same algorithm with different parameters.

Keywords: Multi-Classifier System, ensemble, diversity, heterogeneity.

1 Introduction

Multi-Classifier System (MCS) has gained much attention in pattern recognition, machine learning, and data mining [11, 20]. The key concept of MCS is a combination of different classifiers. It combines predictions from base classifiers into the final predictions [9] and so does an ensemble. MCS is gaining popularity because of the limits of every individual classification algorithm. Techniques proposed to create MCSs or ensembles can be broadly categorized into three categories, according to how base classifiers are constructed. The first category is to use various subsets of training data to train base classifiers, such as boosting [10, 21] and bagging (bootstrap aggregating) [6]. The second category is to use different feature sets to train base classifiers, such as random forest (RF) [7]. Finally, the third category of techniques for the MCS or ensemble construction is to employ different algorithms to build systems composed of heterogeneous classifiers. In fact, these categories represent not only ways to combine classifiers but also ways to achieve diversity among them.

It has been shown that in practice diversity is an important factor that makes them successful. Diversity between two classifiers indicates how different the predictions made by one classifier will be from those made by the other. In this paper, we present

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 923–932, 2009.

theoretical and empirical analyses of classifier combination that is composed of heterogeneous classifiers. We introduce the use of heterogeneous classifiers in MCSs or ensembles, since the classifiers are expected to be uncorrelated and to behave independently from each other. We provide evidence supporting that the diversity among heterogeneous classifiers is better than that among homogeneous ones. We consider four synthetic data sets as well as six real data sets, utilize three of the most important data mining algorithms [25] with or without alternative parameter sets, and evaluate the results using ten well-known diversity measures.

To the best of our knowledge, there exist quite a few papers that theoretically examine the diversity of heterogeneous classifiers and make empirical comparisons of the diversity between homogeneous classifiers to that between heterogeneous classifiers. In our theoretical analysis, we introduce two definitions for diversity based on disagreements. These definitions distinguish themselves from others in the sense that they consider the nature of the underlying algorithm and the training data set as well. Neglecting these two factors, other definitions and measures for diversity failed in investigating effects of the use of heterogeneous classifiers in ensembles. They could not explain the phenomena presented in Section 3. Our contributions are listed below:

1). Two definitions for diversity introduced in this paper could assist researchers in understanding the diversity in MCSs or ensembles more comprehensively.

2) We present a lower bound of the probability that using heterogeneous classifiers would give better diversity when underlying algorithms are different enough.

3) We show that using heterogeneous classifiers consistently provides better diversity regardless of the diversity measures employed in experiments.

4) It is demonstrated that using alternative parameters lead to changes of diversity between homogeneous classifiers but changes are not as significant as those of diversity between heterogeneous classifiers.

The remainder of this paper is organized as follows. In Section 2 we will present the theoretical analysis for diversity between heterogeneous classifiers, while in Section 3 we will report the experiments and results. Next, we will discuss related work in Section 4. Finally, conclusions will be given in Section 5.

2 Diversity of Combinations of Heterogeneous Classifiers

Initially, we introduce notations used in this paper. We use A and S to respectively denote a symmetric classification algorithm and a data set. The i-th sample in a data set S is $z_i = (x_i, y_i)$, $1 \le i \le |S|$, while x_i is a vector of features and y_i is the class label. For convenience and clarity, $X = \{x_i\}$ while $Z = \{z_i\}$ and $Z \subseteq S$. In this paper we consider binary classification problem and hence $y_i = \{-1,+1\}$ and $Y = \{y_i\}$. Moreover, $\hat{y}_i^{(A,S)}$ is the prediction obtained from applying A on S, while I and E_S are respectively the indicator function and the expectation operator with respect to a set S. Below we define two types of diversity. The first one is the intra-algorithm diversity, which is the normalized number of disagreements caused by the nature of the algorithm when dealing with different training sets.

Definition 1. *D is an underlying distribution (which is unknown in practice) and three subsets drawn from it are denoted as S_1, S_2 and S. If S_1 and S_2 are training sets while S is a test set, then the intra-algorithm diversity of a symmetric classification*

algorithm A *is* $\mathrm{E}_S\left[\mathrm{I}(\hat{y}^{(A,S_1)} \neq \hat{y}^{(A,S_2)})\right]$, *an expectation of disagreements with respect to the test set* S.

$\hat{y}_i^{(A,S_1)} \neq \hat{y}_i^{(A,S_2)}$ indicates a disagreement in the prediction of the class label of the i-th sample. For the i-th sample, $\hat{y}_i^{(A,S_1)}$ is the prediction from the classifier created by applying A on S_1, while $\hat{y}_i^{(A,S_2)}$ is that from the classifier created by applying A on S_2. Assume $|S_1| = |S_2| = N$, $|S_1 \cap S_2| = c \cdot |S_1| = c \cdot |S_2| = c \cdot N = n$, where $0 < c < 1$ (which implies $S_1 \cap S_2 \neq \varnothing$). The constant c is used to present the proportion of common samples in both S_1 and S_2. In order not to make the above definition is too strong to be practical, we need restrictions for S, such as $|S_1 \cap S_2| = |S_1 \cap S| = |S_2 \cap S| = c \cdot N = n$. Thus, a simple setting is $S = S_1 \cap S_2$.

Definition 2. *A symmetric classification algorithm* A *is* (α, β)*-stable with respect to the intra-algorithm diversity if* $\Pr\left[\mathrm{E}_{S_1 \cap S_2}[\mathrm{I}(\hat{y}^{(A,S_1)} \neq \hat{y}^{(A,S_2)})] \leq \alpha\right] \geq \beta$ *where* $0 \leq \alpha, \beta \leq 1$.

The (α, β)-stability for the intra-algorithm diversity can be interpreted as follows: α is the upper bound for the normalized number of disagreements under a probability at least β. The second type of diversity is inter-algorithm diversity. It is defined as the normalized number of disagreements due to diverse natures of individual algorithms.

Definition 3. D *is an underlying distribution (which is unknown in practice) and two subsets drawn from it are denoted as* S *and* S'. *If* S *is a training set and* S' *is a test set, the inter-algorithm diversity of two symmetric classification algorithms* A_1 *and* A_2 *is* $\mathrm{E}_{S'}[\mathrm{I}(\hat{y}^{(A_1,S)} \neq \hat{y}^{(A_2,S)})]$ *where* $S' \subseteq S \sim D$ *and* $S' \neq \varnothing$.

$\hat{y}_i^{(A_1,S)} \neq \hat{y}_i^{(A_2,S)}$ indicates a disagreement, with respect to the class label of the i-th sample in S', between the prediction given by a classifier from applying A_1 on S and that given by a classifier from applying A_2 on S. In other words, we apply two algorithms on the whole data set S to build two classifiers, and the diversity between them in this context is defined as the disagreements between these two classifiers for an arbitrary non-empty subset S' of S. What is more, we employ the inter-algorithm diversity to define the differentiability for two classification algorithms.

Definition 4. Two s*ymmetric classification algorithms* A_1 *and* A_2 *are* (δ, γ)*-differentiable if* $\Pr\left[\mathrm{E}_{S'}[\mathrm{I}(\hat{y}^{(A_1,S)} \neq \hat{y}^{(A_2,S)})] \geq \delta\right] \geq \gamma$, *where* $0 \leq \delta, \gamma \leq 1$.

A_1 and A_2 are (δ, γ)-differentiable if δ is the lower bound for the number of disagreements under a probability at least γ. This definition is used to set an assumption that A_1 and A_2 are really different. Afterwards, we consider the combinations of homogeneous and heterogeneous algorithms. We denote a combination of classifiers as $A_a + A_b$ where subscripts a and b indicate the underlying algorithms. Diversities of $A_1 + A_1$ and $A_2 + A_2$ are $\mathrm{E}_{S_1 \cap S_2}\left[\mathrm{I}(\hat{y}^{(A_1,S_1)} \neq \hat{y}^{(A_1,S_2)})\right]$ and $\mathrm{E}_{S_1 \cap S_2}[\mathrm{I}(\hat{y}^{(A_2,S_1)} \neq \hat{y}^{(A_2,S_2)})]$, respectively. The diversity between heterogeneous classifiers can be written as $\mathrm{E}_{S_1 \cap S_2}[\mathrm{I}(\hat{y}^{(A_1,S_1)} \neq \hat{y}^{(A_2,S_2)})]$.

Proposition. *If symmetric classification algorithms* A_1 *and* A_2 *are respectively* (α_1, β_1)*-stable and* (α_2, β_2)*-stable with respect to intra-algorithm diversity, while* A_1 *and* A_2 *are* (δ_2, γ_2)*-differentiable, then the following holds:*

$$\Pr\left[\left(\mathrm{E}_{S_1\cap S_2}[\mathrm{I}(\hat{y}^{(A_1,S_1)}\neq\hat{y}^{(A_2,S_2)})]-\mathrm{E}_{S_1\cap S_2}[\mathrm{I}(\hat{y}^{(A_1,S_1)}\neq\hat{y}^{(A_1,S_2)})]\right)\geq\delta_2-2\cdot\alpha_1\right]\geq\gamma_2\cdot\beta_1 \quad (1)$$

If $\delta_2 > 2\cdot\alpha_1$*, then the diversity of* $A_1 + A_2$ *will be larger than the diversity of* $A_1 + A_1$ *with a probability of at least* $\gamma_2\cdot\beta_1$.

Eq. (1) shows the difference between the diversity from using heterogeneous algorithms and that from using homogeneous ones. First of all, let us focus on the first term $\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)})$, the sum of $\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)}\wedge\hat{y}_i^{(A_1,S_1)}=\hat{y}_i^{(A_1,S_2)})$ and $\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)}\wedge\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})$ since one of them is true. Notice that we consider only binary classification problems. Furthermore, $\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)}$ when $\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)}$ and $\hat{y}_i^{(A_1,S_1)}=\hat{y}_i^{(A_1,S_2)}$, so that $\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)}\wedge\hat{y}_i^{(A_1,S_1)}=\hat{y}_i^{(A_1,S_2)})$ is equal to $\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}=\hat{y}_i^{(A_1,S_2)})$; furthermore, $\hat{y}_i^{(A_1,S_2)}=\hat{y}_i^{(A_2,S_2)}$ when $\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)}$ and $\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)}$, so that $\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)}\wedge\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})$ is equal to $\mathrm{I}(\hat{y}_i^{(A_1,S_2)}=\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})$. So, $\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)})$ gives Eq. (2).

$$\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}=\hat{y}_i^{(A_1,S_2)})+\mathrm{I}(\hat{y}_i^{(A_1,S_2)}=\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)}) \quad (2)$$

Next, $\mathrm{E}_{S_1\cap S_2}[\mathrm{I}(\hat{y}^{(A_1,S_1)}\neq\hat{y}^{(A_2,S_2)})]-\mathrm{E}_{S_1\cap S_2}[\mathrm{I}(\hat{y}^{(A_1,S_1)}\neq\hat{y}^{(A_1,S_2)})]$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)})-\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{l}\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}=\hat{y}_i^{(A_1,S_2)})\\+\mathrm{I}(\hat{y}_i^{(A_1,S_2)}=\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})-\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})\end{array}\right\}\text{ from Eq. (2)}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\{\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})\right\}+\frac{1}{n}\sum_{i=1}^{n}\left\{-2\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})\right\} \quad (3)$$

Under the assumption that A_1 and A_2 are (δ_2, γ_2)-differentiable the probability that the first term in Eq. (3) will be larger than or equal to δ_2 is at least γ_2. Considering the second term in Eq. (3) and $\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})=\{0,1\}$, we have

$$\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})\leq\sum_{i=1}^{n}1\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})$$

We assume A_1 is (α_1, β_1)-stable for intra-algorithm diversity and get the probability that the second term will be larger than or equal to $-2\alpha_1$ is at least β_1.

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_2,S_2)})-\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})-\frac{2}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})\cdot\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})$$

$$\geq\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})-\frac{2}{n}\sum_{i=1}^{n}\mathrm{I}(\hat{y}_i^{(A_1,S_1)}\neq\hat{y}_i^{(A_1,S_2)})\text{ since }\mathrm{I}(\hat{y}_i^{(A_1,S_2)}\neq\hat{y}_i^{(A_2,S_2)})=\{0,1\}$$

$\geq\delta_2-2\cdot\alpha_1$ with probability at least $\gamma_2\cdot\beta_1$, since they are independent events.

Therefore, the diversity given by heterogeneous classifiers is larger than that given by homogeneous classifiers by at least $\delta_2 - 2 \cdot \alpha_1$ with a probability at least $\gamma_2 \cdot \beta_1$. In order to make the above inequality interesting and useful, we need $\delta_2 \geq 2 \cdot \alpha_1$. That is, we need to make sure that the inter-algorithm diversity is at least twice larger than the intra-algorithm diversity of A_1. In summary, according to the above analysis, if we want to obtain better diversity by using heterogeneous classifiers, we should focus on finding an algorithm that is as different from the first algorithm as possible.

3 Experiments and Results

In this paper, we perform experiments with synthetic and real data sets. We exploit two data generators, RDG1 and RandomRBF built in Weka [23], to individually generate two synthetic data sets. We also use six UCI benchmark data sets as real data sets: *breast-w*, *credit-a*, *credit-g*, *diabetes*, *kr-vs-kp*, and *sick* [3, 12].

Diversity measures are proposed to assist in MCS or ensemble design [17]. In [14], authors give a comprehensive summary of diversity measures. Here, we employ ten popular measures to establish the quantitative determination of diversity between classifiers. Diversity measures considered here are summarized below: Q-statistic (*Q*), correlation coefficient (ρ), disagreement measure (*DIS*), double-fault measure (*DF*), entropy (*E*), Kohavi-Wolpert variance (*KW*), interrater agreement (κ), measure of difficulty (θ), generalized diversity (*GD*), and coincident failure diversity (*CFD*).

Furthermore, we consider three disparate algorithms: C4.5 decision tree [19] (named J48 in Weka), Naïve Bayes [13] (NB for short), and nearest neighbor [1] (named IBk in

Table 1. Summary of ten popular diversity measures [16]

Definition	Notations
$Q = (N^{11}N^{00} - N^{01}N^{10}) \cdot (N^{11}N^{00} + N^{01}N^{10})$	N is the number of samples. N^{11}, N^{00}, N^{10}, N^{01} present the number of samples for which both, none, only the1st, only the 2nd classifier(s) made correct prediction, respectively.
$\rho = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11}+N^{10}) \cdot (N^{01}+N^{00}) \cdot (N^{11}+N^{01}) \cdot (N^{10}+N^{00})}}$	
$DIS = (N^{01}+N^{10}) \cdot (N^{11}+N^{10}+N^{01}+N^{00})^{-1}$	
$DF = N^{00} \cdot (N^{11}+N^{10}+N^{01}+N^{00})^{-1}$	
$E = 1/N \sum_{j=1}^{N} \min(l_j, 2-l_j)$	l_j is the number of classifiers that make correct prediction for a sample j. $\overline{p}$ is the average accuracy of all classifiers.
$KW = 1/4N \sum_{j=1}^{N} l_j \cdot (2-l_j)$	
$\kappa = 1-(0.5 \cdot \sum_{j=1}^{N} l_j \cdot (2-l_j)) \cdot (N \cdot \overline{p} \cdot (1-\overline{p}))^{-1}$	
$\theta = \mathrm{var}(X)$ where random variable X denotes the portion of classifiers that make correct prediction for a random sample; That is, $X = \{0, 0.5, 1\}$	
$GD = 1-(P_2(0.5 \cdot P_1 + P_2))^{-1}$	P_i denotes the probability that i randomly selected classifiers will make incorrect prediction
$CFD = P_1 \cdot (1-P_0)^{-1}$ (or 0 if $P_0 = 1$)	

Weka). We select any two of them to create a pair of heterogeneous classifiers so that, for each data set, we set up six experimental sets in each of which we compare diversity of the combination of homogeneous classifiers with that of heterogeneous ones.

As for experiments, we adopt the following procedure to perform experiments, given an input dataset D and algorithms A_1 and A_2: First of all, we randomly draw samples from D without replacement and generate two training datasets. For synthetic datasets, the ratio of a training dataset to the whole dataset (D) is 0.1; for real datasets, it is 0.5. Next, we use one training datasets to create the first classifiers based on an algorithm A_1 , which could be J48, NB, or IBk. We denote this dataset as C_1 . Afterwards, we use the other training dataset to create the second classifier, C_2 , based on the same algorithm A_1 . Next, we create the third classifier, C_3 , by using the second training dataset and another algorithm A_2 (where $A_1 \neq A_2$). Following that, we draw samples from D with replacement and produce ten testing datasets. Then, for each testing dataset, we collect predicted class labels given by C_1 , C_2 , and C_3 as well. Next, we calculate the diversity between C_1 and C_2 (i.e., homogeneous classifiers) and also the diversity between C_1 and C_3 (i.e., heterogeneous classifiers) in ten diversity measures. Finally, we average the diversity values over ten testing datasets.

In the following, we present the results obtained by using the above three classification algorithms with default parameters first and then we present the results obtained by using the algorithms with various parameter sets.

Applying the above procedure to synthetic and real data sets, we collect the diversity values obtained from heterogeneous and homogeneous classifiers. Table A1 and Table A2 show the results for synthetic data sets and real data sets, respectively. For both Table A1 and Table A2, the first column exhibits a group of six experimental sets with respect to data sets. The second column shows the algorithms used in an experimental set. For instance, J48+J48 presents a pair of homogeneous classifiers, while J48+NB presents a pair of heterogeneous classifiers where J48 is the first employed algorithm and NB (i.e., Naïve Bayes) is the second one. For convenience, symbols ↓ and ↑ mean that respectively a higher value and a lower value will give a better diversity with respect to some measure. The results clearly present that using heterogeneous classifiers (shaded rows) leads to better diversity regardless of the diversity measures used in experiments.

According to the above theoretical analysis, the heterogeneity primarily comes from using different algorithms instead of using the same algorithm with different parameters. Here, we provide empirical support for this argument. However, it is impractical and unnecessary to study all possible combinations of parameters. Thus, we consider five quite different parameter sets, as listed in below, to increase the variability of classifiers that are from the same algorithm. Such a selection, or parameter tuning, is a common exercise in data mining.

1) J48: unpruned tree; NB: kernel density estimator; IBK: 3-nearest neighbor
2) J48: minimum 5 instances per leaf; NB: supervised discretization; IBk: 5-nearest neighbor
3) J48: 5-fold reduced error pruning; NB: kernel density estimator; IBk: 5-nearest neighbor, weighted by the inverse of distance
4) J48: confidence threshold 0.2 for pruning, minimum 5 instances per leaf; NB: supervised discretization; IBk: 5-nearest neighbors, weighted by 1-distance
5) J48: unpruned tree, binary splits, minimum 5 instances per leaf; NB: supervised discretization; IBk: 5-nearest neighbors, hold-one-out evaluation for training, minimizing mean squared error

We applied algorithms with these parameter sets on all data sets we mentioned earlier. However, due to the limitation of space, we do not report all results; rather, we present here the results for a synthetic data set and a real data set. Table A3 and Table A4 give the results of applying such pairs of homogeneous classifiers to a synthetic data set and the real data set *diabetes*, respectively. The first column shows parameter sets and the second column shows experiments for homogeneous and heterogeneous classifiers. The superscript asterisk means that, the pair of homogeneous classifiers is constructed with different parameters while one of them (the second one) comes with alternative parameters corresponding. The results presented in Tables A3 and A4 are not as optimistic as those reported in Tables A1 and A2. For the combination of homogeneous classifiers, using different parameters indeed gives better diversity. Changing parameters does not mean changing the nature of an algorithm but the way the algorithm searches the hypothesis space, if we interpret it in the classical language of machine learning. However, the diversity among classifiers that are based on the same algorithm but come with different parameters would not be good enough to differentiate them. From these results, in the setting considered here, the primary source of heterogeneity is the mix of different algorithms rather than the use of the same algorithm with different parameters. Nevertheless, the conclusion by no means indicates that employing different parameters has no effect on the diversity. It will be interesting to study the theoretical relationship between diversity and this factor.

4 Related Work

The study of diversity in ensemble has gained increasing attention, even though in theory there is no strong connection between diversity and the overall accuracy [5, 8, 14, 15, 16, 18, 20, 23].In [18] authors indicate that, in general, diversity compensates for errors made by individual classifiers. However, diversity itself is not a strong predictor of the overall accuracy of an ensemble [17]. In [15] authors discuss the relationship between diversity and accuracy, while in [23] it is indicated that an effective ensemble requires each of individual classifiers to offer high accuracy and to generate diverse errors. Additionally, in [22], authors demonstrate that boosting requires stronger diversity than does bagging while bagging does not depend only on diversity, and they argue that diversity depends on the size of training data set. Moreover, in [2] authors consider using different feature sets argue that using different feature sets is the only way to achieve diversity in a system of homogeneous classifiers. However, the argument is not necessarily true because using homogeneous classifiers with different parameters would lead to the change of diversity, as we can see in tables. Furthermore, we connect heterogeneity to diversity without considering the effects of using different feature sets. In [4] authors study the combination of heterogeneous classifiers with the focus on some combination methods. Nevertheless, neither a theoretical analysis nor an empirical investigation of the source of heterogeneity is performed in the paper.

5 Conclusions

This paper studied theoretically and empirically the relationship between heterogeneity and diversity. We performed a rich set of experiments to provide empirical evidence.

To evaluate the results, we considered four synthetic data sets as well as six real benchmark data sets, utilized three classification algorithms without and with five different parameter sets, and employed ten popular diversity measures. Consequently, we built a foundation for the use of heterogeneous classifiers in MCSs or ensembles. This is particularly essential because there are quite a few papers theoretically examining the diversity of classifier combinations and, at the same time, empirically comparing the diversity of the combination of homogeneous classifiers with that of heterogeneous classifiers. Two important observations in this paper will make substantial contributions to the future MCS or ensemble design. First, the diversity among heterogeneous classifiers is higher than that among homogeneous ones. Second, the heterogeneity mainly results from using different classification algorithms instead of using the same algorithm with different parameters. Future work includes the study of the relationship between heterogeneity and accuracy.

References

1. Aha, D., Kibler, D.: Instance-based learning algorithms. Machine Learning 6(1), 37–66 (1991)
2. Alkoot, F.M., Kittler, J.: Multiple expert system design by combined feature selection and probability level fusion. In: Proc. of the 3rd International Conference on Information Fusion, vol. 2, pp. THC5/9–THC516 (2000)
3. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
4. Bahler, D., Navarro, L.: Methods for Combining Heterogeneous Sets of Classifiers. In: The 17th National Conference on Artificial Intelligence, Workshop on New Research Problems for Machine Learning (2000)
5. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: A New Ensemble Diversity Measure Applied to Thinning Ensembles. In: International Workshop on Multiple Classifier Systems, pp. 306–316 (2003)
6. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
7. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
8. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: a survey and categorization. Information Fusion 6(1), 5–20 (2005)
9. Dietterich, T.G.: Ensemble Methods in Machine Learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
10. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: Proc. of the 13th International Conference on Machine Learning, pp. 148–156 (1996)
11. Ghosh, J.: Multiclassifier Systems: Back to the Future. In: Roli, F., Kittler, J. (eds.) MCS 2002. LNCS, vol. 2364, pp. 1–15. Springer, Heidelberg (2002)
12. Hettich, S., Bay, S.D.: The UCI KDD Archive. University of California, Department of Information and Computer Science, Irvine, CA (1999), http://kdd.ics.uci.edu
13. John, G.H., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In: The 11th Conference on Uncertainty in Artificial Intelligence, pp. 338–345 (1995)
14. Kuncheva, L.I., Whitaker, C.J.: Ten measures of diversity in classifier ensembles: limits for two classifiers. In: A DERA/IEE Workshop on Intelligent Sensor Processing, pp. 10/1–10/10 (2001)

15. Kuncheva, L.I., Skurichina, M., Duin, R.P.W.: An experimental study on diversity for bagging and boosting with linear classifiers. Information Fusion 3(4), 245–258 (2002)
16. Kuncheva, L.I., Whitaker, C.J.: Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51(2), 181–207 (2003)
17. Kuncheva, L.I.: That elusive diversity in classifier ensembles. In: Proc. of Iberian Conference on Pattern Recognition and Image Analysis, pp. 1126–1138 (2003)
18. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of AI Research 11, 169–198 (1999)
19. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
20. Ranawana, R.: Multi-Classifier Systems - Review and a Roadmap for Developers. International Journal of Hybrid Intelligent Systems 3(1), 35–61 (2006)
21. Schapire, R.E.: The boosting approach to machine learning: An overview. In: MSRI Workshop on Nonlinear Estimation and Classification (2002)
22. Skurichina, M., Kuncheva, L., Duin, R.P.: Bagging and Boosting for the Nearest Mean Classifier: Effects of Sample Size on Diversity and Accuracy. In: Roli, F., Kittler, J. (eds.) MCS 2002. LNCS, vol. 2364, pp. 62–71. Springer, Heidelberg (2002)
23. Valentini, G., Masulli, F.: Ensembles of Learning Machines. In: Marinaro, M., Tagliaferri, R. (eds.) WIRN 2002. LNCS, vol. 2486, pp. 3–22. Springer, Heidelberg (2002)
24. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
25. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 Algorithms in Data Mining. Knowledge and Information Systems 14(1), 1–37 (2008)

Appendix

Table A1. Average diversity values obtained from combinations of heterogeneous classifiers (shaded rows) and those of homogeneous classifiers on four synthetic data sets

Exp.	Algo.	$Q\downarrow$	$\rho\downarrow$	$DIS\uparrow$	$DF\downarrow$	$E\uparrow$	$KW\uparrow$	$\kappa\downarrow$	$\theta\downarrow$	$GD\uparrow$	$CFD\uparrow$
Set 1	J48+J48	1	1	0	0.225	0	0	1	0.1175	0	0
	J48+NB	0.38	0.1375	0.3375	0.1125	0.3375	0.085	0.325	0.05	0.64	0.775
Set 2	J48+J48	1	1	0	0.225	0	0	1	0.1175	0	0
	J48+IBk	0.705	0.2575	0.2	0.0625	0.2	0.05	0.595	0.0975	0.6625	0.7925
Set 3	NB+NB	1	1	0	0.3425	0	0	1	0.085	0	0
	NB+J48	0.38	0.1375	0.3375	0.1125	0.3375	0.085	0.325	0.05	0.64	0.775
Set 4	NB+NB	1	1	0	0.3425	0	0	1	0.085	0	0
	NB+IBk	0.32	0.0875	0.3375	0.0525	0.3375	0.085	0.32	0.065	0.79	0.88
Set 5	IBk+IBk	1	1	0	0.1025	0	0	1	0.1625	0	0
	IBk+J48	0.705	0.2575	0.2	0.0625	0.2	0.05	0.595	0.0975	0.6625	0.7925
Set 6	IBk+IBk	1	1	0	0.1025	0	0	1	0.1625	0	0
	IBk+NB	0.32	0.0875	0.3375	0.0525	0.3375	0.085	0.32	0.065	0.79	0.88

Table A2. Average diversity values obtained from combinations of heterogeneous classifiers (shaded rows) and those of homogeneous classifiers on six real data sets.

Exp.	Algo.	$Q\downarrow$	$\rho\downarrow$	$DIS\uparrow$	$DF\downarrow$	$E\uparrow$	$KW\uparrow$	$\kappa\downarrow$	$\theta\downarrow$	$GD\uparrow$	$CFD\uparrow$
Set 1	J48+J48	1	1	0	0.095	0	0	1	0.1717	0	0
	J48+NB	0.8133	0.32	0.1367	0.0533	0.1367	0.0333	0.73	0.1217	0.6483	0.7633
Set 2	J48+J48	1	1	0	0.095	0	0	1	0.1717	0	0
	J48+IBk	0.8583	0.3117	0.0983	0.0333	0.0983	0.0233	0.8033	0.1483	0.6633	0.79
Set 3	NB+NB	1	1	0	0.15	0	0	1	0.1417	0	0
	NB+J48	0.8133	0.32	0.1367	0.0533	0.1367	0.0333	0.73	0.1217	0.6483	0.7633
Set 4	NB+NB	1	1	0	0.15	0	0	1	0.1417	0	0
	NB+IBk	0.665	0.2	0.1517	0.035	0.1517	0.0367	0.6983	0.1233	0.75	0.8517
Set 5	IBk+IBk	1	1	0	0.0683	0	0	1	0.18	0	0
	IBk+J48	0.8583	0.3117	0.0983	0.0333	0.0983	0.0233	0.8033	0.1483	0.6633	0.79
Set 6	IBk+IBk	1	1	0	0.0683	0	0	1	0.18	0	0
	IBk+NB	0.665	0.2	0.1517	0.035	0.1517	0.0367	0.6983	0.1233	0.75	0.8517

Table A3. Average diversity values from combinations of homogeneous (HO*) and heterogeneous (HE) classifiers (HE) on a synthetic data set

Para. sets	Exp.	$Q\downarrow$	$\rho\downarrow$	$DIS\uparrow$	$DF\downarrow$	$E\uparrow$	$KW\uparrow$	$\kappa\downarrow$	$\theta\downarrow$	$GD\uparrow$	$CFD\uparrow$
1	HO*	1	0.8567	0.0733	0.3267	0.0733	0.02	0.8533	0.06	0.13	0.1867
	HE	0.34	0.16	0.4	0.1267	0.4	0.1	0.2	0.0333	0.6033	0.7467
2	HO*	0.9733	0.7367	0.13	0.31	0.13	0.03	0.74	0.0433	0.2133	0.3133
	HE	0.34	0.16	0.4	0.1267	0.4	0.1	0.2	0.0333	0.6033	0.7467
3	HO*	0.9267	0.8067	0.0833	0.29	0.0833	0.02	0.8333	0.0667	0.13	0.1967
	HE	0.34	0.16	0.4	0.1267	0.4	0.1	0.2	0.0333	0.6033	0.7467
4	HO*	0.9267	0.8067	0.0833	0.29	0.0833	0.02	0.8333	0.0667	0.13	0.1967
	HE	0.34	0.16	0.4	0.1267	0.4	0.1	0.2	0.0333	0.6033	0.7467
5	HO*	0.9767	0.89	0.05	0.3067	0.05	0.0133	0.9	0.0767	0.0733	0.12
	HE	0.36	0.1667	0.3883	0.125	0.3883	0.0967	0.2233	0.0333	0.6017	0.745

Table A4. Average diversity values from combinations of homogeneous (HO*) and heterogeneous (HE) classifiers (HE) on the real data set *diabetes*

Para. sets	Exp.	$Q\downarrow$	$\rho\downarrow$	$DIS\uparrow$	$DF\downarrow$	$E\uparrow$	$KW\uparrow$	$\kappa\downarrow$	$\theta\downarrow$	$GD\uparrow$	$CFD\uparrow$
1	HO*	0.97	0.79	0.0733	0.1833	0.0733	0.02	0.8533	0.09	0.1767	0.27
	HE	0.7533	0.3833	0.2067	0.1067	0.2067	0.0467	0.5967	0.0667	0.51	0.67
2	HO*	0.9067	0.5867	0.1467	0.15	0.1467	0.0367	0.71	0.07	0.3333	0.4867
	HE	0.7533	0.3833	0.2067	0.1067	0.2067	0.0467	0.5967	0.0667	0.51	0.67
3	HO*	0.8767	0.6033	0.1267	0.1433	0.1267	0.0333	0.74	0.0833	0.3167	0.4633
	HE	0.7533	0.3833	0.2067	0.1067	0.2067	0.0467	0.5967	0.0667	0.51	0.67
4	HO*	0.9067	0.5867	0.1467	0.15	0.1467	0.0367	0.71	0.07	0.3333	0.4867
	HE	0.7533	0.3833	0.2067	0.1067	0.2067	0.0467	0.5967	0.0667	0.51	0.67
5	HO*	0.9367	0.63	0.13	0.1567	0.13	0.0333	0.74	0.0733	0.3033	0.4567
	HE	0.7533	0.3833	0.2067	0.1067	0.2067	0.0467	0.5967	0.0667	0.51	0.67