Mach Learn (2006) 65:247–271
DOI 10.1007/s10994-006-9449-2

An analysis of diversity measures

E. K. Tang · P. N. Suganthan · X. Yao

Received: 5 January 2005 / Revised: 6 April 2006 / Accepted: 30 May 2006 / Published online: 14 July 2006

Abstract Diversity among the base classifiers is deemed to be important when constructing a classifier ensemble. Numerous algorithms have been proposed to construct a good classifier ensemble by seeking both the accuracy of the base classifiers and the diversity among them. However, there is no generally accepted definition of diversity, and measuring the diversity explicitly is very difficult. Although researchers have designed several experimental studies to compare different diversity measures, usually confusing results were observed. In this paper, we present a theoretical analysis on six existing diversity measures (namely disagreement measure, double fault measure, KW variance, inter-rater agreement, generalized diversity and measure of difficulty), show underlying relationships between them, and relate them to the concept of margin, which is more explicitly related to the success of ensemble learning algorithms. We illustrate why confusing experimental results were observed and show that the discussed diversity measures are naturally ineffective. Our analysis provides a deeper understanding of the concept of diversity, and hence can help design better ensemble learning algorithms.

Keywords Classifier ensemble · Diversity measures · Margin distribution · Majority vote · Disagreement measure · Double fault measure · KW variance · Interrater agreement · Generalized diversity · Measure of difficulty · Entropy measure · Coincident failure diversity

Editor: Tom Fawcett

E. K. Tang · P. N. Suganthan
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798
e-mail: tangke@pmail.ntu.edu.sg

P. N. Suganthan (✉)
e-mail: epnsugan@ntu.edu.sg

X. Yao
School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK
e-mail: X.Yao@cs.bham.ac.uk

1. Introduction

Also known as committees of learners, mixtures of experts, multiple classifier systems, and classifier ensembles, ensemble learning has been well established as a research area in the past decades. Some well-known algorithms in this field, such as bagging (Breiman, 1996) and boosting (Freund, 1995; Freund & Schapire, 1996), have been proven to be effective in solving pattern classification tasks. Each classifier in the ensemble is referred as a base classifier. Intuitively speaking, the key to the success of a classifier ensemble is that the base classifiers perform diversely. Empirical results have illustrated that there exists positive correlation between accuracy of the ensemble and diversity among the base classifiers (Dietterich, 2000; Kuncheva & Whitaker, 2003a). Further, most of the existing ensemble learning algorithms (Brieman, 1996; Ho, 1998; Schapire, 1999; Suganthan, 1999; Liu et al., 2000; Atukorale et al., 2003) can be interpreted as building diverse base classifiers implicitly.

In an ensemble learning algorithm, if the term "diversity" is defined explicitly and optimized, we say the algorithm seeks diversity explicitly. Otherwise the algorithm seeks diversity implicitly. Since many ensemble algorithms have been successfully proposed by seeking diversity implicitly, it is natural to consider whether we can perform better by seeking the diversity explicitly. However, despite the popularity of the term diversity, there is no single definition and measure of it. Although several measures have been proposed to represent the diversity and are optimized explicitly in different ensemble learning algorithms, none of these measures is proven superior to the others and why these diversity measures are useful is still unclear. Solid empirical as well as theoretical validation of the explicitly diversity-driven algorithms is absent from this field. In particular, three main questions in the analysis of diversity have arisen from several previous empirical studies (Bauer & Kohavi, 1999; Dietterich, 2000; Kuncheva & Whitaker, 2003a) in relation to designing ensemble learning algorithms:

1. When one seeks a set of diverse and accurate base classifiers, what cost function is optimized? Does optimizing the cost function guarantee good generalization performance?
2. Is there a trade-off between diversity and accuracy of the base classifiers? In other words, do we have to sacrifice the accuracy of some base classifiers in order to increase the diversity?
3. How to make use of the existing diversity measures for designing good classifier ensembles? Besides the existing diversity measures, is a new precise diversity measure necessary for designing ensemble learning algorithms?

Although all of the three questions are important, none of them has been thoroughly answered, while several contradictory conclusions have been drawn from experimental studies. On the other hand, in spite of the incomplete understanding of the concept of diversity, more complete theories have been proposed to explain the success of classifier ensembles. For example, Schapire et al. (1998) introduced the concept of margin to analyze the behavior of boosting type algorithms. This concept was then easily generalized to analyze other classes of ensemble learning algorithms (Mason et al., 2000; Breiman, 2001). Further, bias-variance decomposition is also employed to explain the success of classifier ensembles (Liu & Yao, 1997).

Motivated by both the previous experimental and theoretical studies, we present an in-depth analysis of some existing diversity measures in this paper. We analyze the underlying relationships between these measures and the margin of a classifier ensemble, which put the study of diversity measures into the context of so-called large margin classifiers. Following

this framework, the three questions presented above are answered by both theoretical and experimental analysis.

The remainder of this paper is organized as follows. We first introduce some relevant concepts of classifier ensembles, diversity measures and the margin. In Section 3, we analyze theoretically and experimentally the relationships among diversity measures, average classification accuracy of base classifiers and margin of the ensemble. We further discuss the applications of diversity measures in an ensemble learning algorithm in Section 4. Conclusions are presented in Section 5.

2. Related background

Let a labeled training set be $\mathbf{Tr} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, where y_i is the class label of $\mathbf{x}_i$. The base classifiers $\mathbf{H} = \{h_1, h_2, \ldots, h_L\}$ of an ensemble are trained on the training set, and the output of a base classifier h_j on sample $\mathbf{x}_i$ is $h_j(\mathbf{x}_i)$. Given this set of base classifiers, together with a corresponding set of weights $\mathbf{w} = [w_1, \ldots, w_L]^T$, where $w_j \geq 0$ and $\sum w_j = 1$, the ensemble classifies the samples by taking a weighted vote among the base classifiers and choosing the class label that receives the largest weighted vote. For every $1 \leq i \leq N$ and $1 \leq j \leq L$, the oracle output matrix $\mathbf{O}$ of an ensemble is defined as:

$O_{ij} = 1$ if training sample $\mathbf{x}_i$ is classified correctly by base classifier h_j
$O_{ij} = -1$ otherwise

Hence, the oracle output matrix is an N-by-L matrix whose elements take either 1 or –1. We focus our investigation on the oracle output of the classifiers because:

1. Many other authors discussed the diversity in terms of oracle outputs, and most of the diversity measures are defined based on oracle outputs. Actually, all of the ten diversity measures summarized by Kuncheva and Whitaker (2003a) are based on oracle outputs and the majority vote rule.
2. The oracle output incorporates no *a priori* knowledge of the data and makes no assumption on what the base classifier is. It only concerns whether a sample is classified correctly or not. Hence, the oracle output provides a general model for analyzing a classifier ensemble, and conclusions drawn on this model can be easily generalized to various ensemble learning methods.

The main notations used in this paper are summarized as follows:

L: total number of base classifiers
N: total number of training samples
m_i: margin of an ensemble on the training sample $\mathbf{x}_i$
P: average classification accuracy of the base classifiers on the training data
p_j: classification accuracy of the base classifier h_j
l_i: product of L and sum of the weights of the base classifiers that classify the training sample $\mathbf{x}_i$ incorrectly, $l_i = L \sum_{O_{ij}=-1} w_j$
O_{ij}: oracle output of the classifier h_j on the training sample $\mathbf{x}_i$
div: diversity among the base classifiers in the ensemble

The definitions above directly yield the following equations:

$$P = \sum_{j=1}^{L} w_j p_j, \tag{1}$$

$$P = 1 - \frac{\sum_{i=1}^{N} l_i}{NL}, \tag{2}$$

$$p_j = \frac{1}{2} + \frac{\sum_{i=1}^{N} O_{ij}}{2N}. \tag{3}$$

When all base classifiers of the ensemble are uniformly weighted (i.e. $w_j = 1/L$ for $j =$ 1 to L), we have the majority vote rule. Since majority vote is simple and is employed in most works concerned with diversity measures, we will first present our analysis based on it and then generalize our conclusions to the non-uniformly weighted case. For the uniformly weighted case, we simply replace the weights $1/L$ with 1 for the definition of l_i. Hence l_i becomes the number of base classifiers that classify the training sample $\mathbf{x}_i$ incorrectly. This small modification will not influence our analysis. In the subsequent sections, unless we refer explicitly to the non-uniform weights, all definitions and derivations are based on majority vote rule.

2.1. The diversity measures

Since a unique definition of diversity does not exist, we consider six diversity measures in this work. These measures were proposed by different researchers independently. When we mention diversity, we refer to the diversity that is defined by these measures.

2.1.1. The disagreement measure

In 1996, Skalak (1996) proposed the disagreement measure to evaluate the diversity between two base classifiers. Ho (1998) also employed the disagreement to measure diversity in a decision forest. This measure is defined based on the intuition that two diverse classifiers perform differently on the same training data. Given two base classifiers h_j and h_k, let $n(a, b)$ be the number of training samples on which the oracle output of h_j and h_k is a and b respectively. The diversity between the two base classifiers is measured by:

$$dis_{j,k} = \frac{n(1,-1) + n(-1,1)}{n(1,1) + n(-1,1) + n(1,-1) + n(-1,-1)}. \tag{4}$$

Diversity within the whole set of base classifiers is then calculated by averaging over all pairs of base classifiers:

$$dis = \frac{2}{L(L-1)} \sum_{j=1}^{L} \sum_{k=j+1}^{L} dis_{j,k}. \tag{5}$$

Since for any pair of base classifiers: $n(1,1)+n(1,-1)+n(-1,1)+n(-1,-1)=N$, we can get:

$$dis = \frac{2}{NL(L-1)} \sum_{j=1}^{L} \sum_{k=j+1}^{L} (n_{j,k}(1,-1) + n_{j,k}(-1,1)). \tag{6}$$

The diversity increases with the value of the disagreement measure.

2.1.2. The double-fault measure

Giacinto and Roli (2001) proposed the double-fault measure to select classifiers that are least related from a pool of classifiers. The double-fault measure between a pair of base classifiers is calculated by:

$$DF_{j,k} = \frac{n(-1,-1)}{n(1,1)+n(-1,1)+n(1,-1)+n(-1,-1)}. \tag{7}$$

This measure also arose from the intuition that two classifiers should perform differently to be diverse. Giacinto and Roli claimed that the more different two classifiers are, the fewer the coincident errors between them. Same as the disagreement measure, the diversity within the whole set of base classifiers is calculated as follows:

$$DF = \frac{2}{NL(L-1)} \sum_{j=1}^{L} \sum_{k=j+1}^{L} n_{j,k}(-1,-1). \tag{8}$$

The diversity decreases when the value of the double-fault measure increases.

2.1.3. Kohavi-Wolpert variance

The Kohavi-Wolpert variance was proposed by Kohavi and Wolpert (1996) in their decomposition formula of the classification error of a classifier. This measure originated from the bias-variance decomposition of the error of a classifier. The original expression of the variability of the predicted class label y for a sample $\mathbf{x}$ is

$$variance_{\mathbf{x}} = \frac{1}{2}\left(1 - \sum_{i=1}^{C} P(y=\omega_i|\mathbf{x})^2\right) \tag{9}$$

where C is the number of classes. Since $C=2$ in the case of oracle output, and $P(y=1\,|\,\mathbf{x}) + P(y=-1\,|\,\mathbf{x}) = 1$, we can get

$$\begin{aligned} variance_{\mathbf{x}} &= \frac{1}{2}(1 - P(y=1\,|\,\mathbf{x})^2 - P(y=-1\,|\,\mathbf{x})^2) = P(y=1\,|\,\mathbf{x})P(y=-1\,|\,\mathbf{x}) \\ &= P(O=1\,|\,\mathbf{x})P(O=-1\,|\,\mathbf{x}) \end{aligned}$$

As the term $P(O = -1 \mid \mathbf{x})$ can be estimated by $P(O = -1 \mid \mathbf{x}) = \frac{l_i}{L}$, Kuncheva and Whitaker (2003a) presented a modified version of Eq. (9) to measure the diversity of an ensemble:

$$KW = \frac{1}{NL^2} \sum_{i=1}^{N} l_i(L - l_i). \tag{10}$$

The diversity increases with values increasing of the *KW* variance.

2.1.4. Measurement of inter-rater agreement

This measure is developed as a measure of inter-rater (inter-classifier) reliability (Fless, 1981), called k. It can be used to measure the level of agreement within a set of classifiers, hence it is also based on the assumption that a set of classifiers should disagree with one another to be diverse. The diversity decreases when the value of k increases. The k is calculated by:

$$k = 1 - \frac{\sum_{i=1}^{N} (L - l_i) l_i}{NL(L-1)P(1-P)}. \tag{11}$$

2.1.5. Generalized diversity

This measure is proposed by Partidge and Krzanowski (1997). The heuristic behind this measure is similar to that of the Double-Fault measure. Given two classifiers, Partidge and Krzanowski argued that maximum diversity is achieved when failure of one classifier is accompanied by correct classification by the other classifier and minimum diversity occurs when two classifiers fail together. Therefore, for a sample $\mathbf{x}_i$ that is randomly drawn from the training set, let T_j denote the probability that $l_i = j$, the generalized diversity is defined as:

$$GD = 1 - \frac{\sum_{j=1}^{L} \frac{j(j-1)}{L(L-1)} T_j}{\sum_{j=1}^{L} \frac{j}{L} T_j}. \tag{12}$$

The diversity increases with increasing values of the generalized diversity.

2.1.6. The measure of "difficulty"

This measure comes from the study of Hansen and Salamon (1990). Defining a discrete random variable V, $V_i = (L - l_i)/L$ for a sample $\mathbf{x}_i$ that is randomly drawn from the training set, the measure of difficulty was defined as the variance of V over the whole training set.

$$diff = var(V_i). \tag{13}$$

The diversity increases with decreasing values of the measure of difficulty. The intuition of this measure can be explained as: A diverse classifier ensemble has a smaller value for this measure since every training sample can at least be classified correctly by a portion of all the base classifiers, which is likely to result in lower variance of V.

Table 1 Summary of diversity measures

	dis	*DF*	*KW*	*k*	*GD*	*diff*
div	+	−	+	−	+	−

Table 1 shows a summary of the six measures. The "+" means that diversity is greater when the measure is larger, and the "−" means that diversity is greater when the measure is smaller.

2.2. The margin of ensembles

To explain the remarkable success of the boosting algorithm, Schapire et al. (1998) introduced the concept of margin into the area of ensemble learning. Let $v_{i,C}$ be the total vote that the weighted ensemble casts for label C on sample $\mathbf{x}_i$. The margin of the ensemble on this sample is defined as $m_i = v_{i,y_j} - \sum_{c \neq y_j} v_{i,c}$. Given a set of base classifiers and the weights $\mathbf{w} = [w_1, \ldots, w_L]^T$, where $w_j \geq 0$ and $\sum w_j = 1$, margin of the ensemble can be calculated by

$$m_i = \sum_{j=1}^{L} w_j O_{ij}. \tag{14}$$

Since boosting type algorithms construct an ensemble with respect to the vote rule, the margin concept was easily generalized to all the ensembles that employ the vote rule as the combination method (Schapire et al., 1998). Several extensive studies have shown that the generalization performance of an ensemble is related to the distribution of its margins on the training samples. Schapire et al. proved that achieving a larger margin on the training set results in an improved bound on the generalization error of the ensemble. They also proposed the upper bound explicitly as the sum of a function of distribution of the margins and a complexity penalty term (Schapire et al., 1998). Rätsch et al. (2001) analyzed the margins by focusing on the "minimum margin" of the ensemble on training samples. Employing boosting as an example, they explained the good generalization performance of an ensemble in terms of the minimum margin that can be achieved by it: the ensemble with the largest minimum margin will have the best generalization error bound[1] (Vapnik, 1995; Schapire et al., 1998). Since the generalization error itself is actually immeasurable, this relationship provides us an approach to analyze the relationship between diversity and generalization performance.

3. Analysis of diversity measures

3.1. Relationship between diversity and margins of an ensemble (majority vote case)

If we regard the generalization performance of an ensemble as a function F that is parameterized by the average classification accuracy P of the base classifiers and the diversity among the base classifiers (*div*), our analysis can be formulated as follows: We analyze how F changes with respect to *div* if P is fixed, and how P and *div* interact with each other to influence F.

[1]When the data is noisy, the ensemble that is constructed by maximizing the margin may over-fit, which means the generalization performance of the ensemble may decrease when the minimum margin is maximized. Some discussion is presented in Section 3.

Aiming to maximize the diversity between base classifiers, we begin with answering the first question that is mentioned in Section 1, the cost function for the six diversity measures can be reformulated as:

$$\text{The disagreement measure:} \quad dis = \frac{2L(1-P)}{(L-1)} - \frac{2}{NL(L-1)}\sum_{i=1}^{N} l_i^2 \tag{15}$$

$$\text{The double-fault measure:} \quad DF = \frac{1}{NL(L-1)}\sum_{i=1}^{N} l_i^2 - \frac{1-P}{L-1} \tag{16}$$

$$\text{The Kohavi-Wolpert variance:} \quad KW = 1 - P - \frac{1}{NL^2}\sum_{i=1}^{N} l_i^2 \tag{17}$$

$$\text{The measurement of inter-rater agreement:} \quad k = \frac{LP - P - L}{LP - P} + \frac{\sum_{i=1}^{N} l_i^2}{NL(L-1)P(1-P)} \tag{18}$$

$$\text{The generalized diversity:} \quad GD = \frac{L}{L-1} - \frac{\sum_{i=1}^{N} l_i^2}{NL(L-1)(1-P)} \tag{19}$$

$$\text{The measure of "Difficulty":} \quad diff = \frac{1}{NL}\sum_{i=1}^{N} l_i^2 - L\,(1-P)^2 \tag{20}$$

It can be observed from expressions (15)–(20) that all these cost functions contain the terms P and $\sum_{i=1}^{N} l_i^2$. Based on this observation, we propose the statement:

Lemma 1. *If we regard the average classification accuracy P of the base classifiers as a constant, the diversity (div) is maximized only when all the training samples are classified correctly by the same number of base classifiers, which means*:

$$l_i = L\,(1-P) \quad \forall i. \tag{21}$$

We call Eq. (21) the *uniformity condition* for maximizing the diversity. Appendix A provides detailed derivations of Eqs. (15)–(20) and proof of Lemma 1.

In the uniformly weighted case the margin of an ensemble on sample $\mathbf{x}_i$ is calculated by equations:

$$m_i = \frac{1}{L}\sum_{j=1}^{L} O_{ij} \tag{22}$$

and

$$m_i = \frac{L - 2l_i}{L}. \tag{23}$$

then

$$\sum_{i=1}^{N} m_i = \frac{1}{L}\sum_{i=1}^{N}(L - 2l_i)$$
$$= N(2P - 1)\,. \tag{24}$$

Since $\min(m_i) \leq \frac{1}{N}\sum_{i=1}^{N} m_i$,

$$\min(m_i) \leq 2P - 1 \tag{25}$$

As mentioned in Section 2.2, the best generalization error bound of an ensemble can be achieved by maximizing the minimum margin of the ensemble on the training samples. Hence, the best generalization error bound of the ensemble is achieved when $\min(m_i) = 2P - 1$. It is obvious that the equality in Eq. (25) holds only when:

$$m_i = 2P - 1 \qquad \forall i, \tag{26}$$

$$which\ means: \qquad l_i = L(1 - P) \quad \forall i. \tag{21}$$

Therefore, for the six diversity measures discussed in this work, the answer to the first question posed in the Section 1 can be summarized by Eqs. (15)–(20) and the Lemma below:

Lemma 2. *If P is regarded as a constant and the maximum diversity is achievable, maximizing the diversity among the base classifiers is equivalent to maximizing the minimum margin of the ensemble on the training samples.*

As a result, seeking diversity in an ensemble can be viewed as an implicit way to maximize the minimum margin of the ensemble. The key observations presented in previous experimental studies can be explained by this relationship.

3.2. Ineffectiveness of the diversity measures

In addition to theoretical derivations, we conducted several experiments to illustrate relationships among the average accuracy P, the diversity and the minimum margin of an ensemble. In the first experiment, setting the number of samples $N = 100$, the number of base classifiers $L \in \{15, 150, 1500\}$ and the average accuracy $P = 0.7$, we randomly generate 10000 pseudo oracle output matrices. These matrices are used to represent a set of classifier ensembles. The corresponding diversity measures and the minimum margins are calculated and relationships among them are plotted. The maximum minimum margin in this case is 0.4. For each diversity measure, the optimized value is also presented in the figures. Similar plots are observed for all the three values of L, but we only plot those figures corresponding to $L = 150$ in Figs. 1.1–1.6 to save space.

In the past, experimental study has revealed that large diversity does not always correspond to good generalization performance even when P is fixed (Kuncheva & Whitaker, 2003a). That is why usefulness of diversity measures was questioned. From the previous subsection and Figs. 1.1–1.6, two reasons of this discrepancy can be summarized:

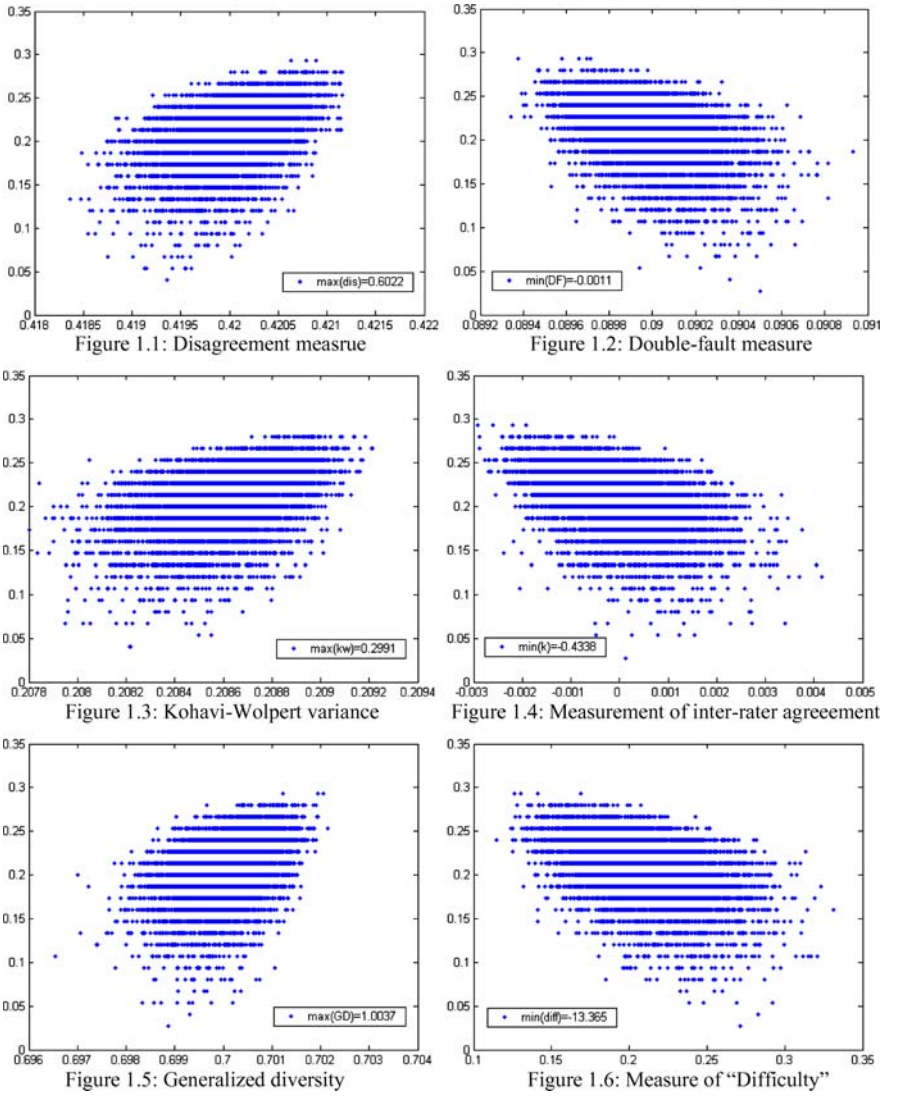

Figure 1.1: Disagreement measrue

Figure 1.2: Double-fault measure

Figure 1.3: Kohavi-Wolpert variance

Figure 1.4: Measurement of inter-rater agreeement

Figure 1.5: Generalized diversity

Figure 1.6: Measure of "Difficulty"

Fig. 1 The figures show the relationships between diversity measures and the minimum margin. $L = 150$, horizontal axis represents the diversity measures and vertical axis represents the minimum margin

1. *For a given P, the maximum diversity is usually not achievable.*

According to the definition, l_i's take discontinuous values, while $L(1-P)$ is a continuous variable since P is continuous. Hence, Eq. (21) cannot be satisfied in general. When we attempt to seek diversity, we usually achieve only *large* diversity but not the *maximum* diversity.

2. *The minimum margin of an ensemble is not monotonically increasing with respect to diversity*.

From Figs. 1.1–1.6, we observed that the minimum margin of a classifier ensemble is not monotonically increasing with the diversity, albeit positive correlation is as obvious as in the literature. Therefore, enlarging diversity is different from enlarging the minimum margin.

When P is a constant, we can find from Eqs. (15)–(20) that the smaller the term $\sum_{i=1}^{N} l_i^2$, the larger the diversity. From Eq. (23), we can also find that the minimum margin is determined by the maximum l_i. The smaller the $\max(l_i)$, the larger the minimum margin. Since a smaller $\sum_{i=1}^{N} l_i^2$ does not always mean a smaller $\max(l_i)$, a larger value for the diversity does not guarantee a larger value for the minimum margin.

The two above-mentioned reasons result in the fact that large diversity may not consistently correspond to a better generalization performance. This is a main drawback of using diversity measures in the implementation of an ensemble learning algorithm.

The first experiment only analyzed the diversity measures under the assumption that P is a constant. However, P and *div* actually interact with each other to influence the performance of an ensemble. By taking a closer look at Eqs. (21) and (25), one finds that P determines the upper bound of the minimum margin, while distribution of l_i over the training set determines difference between the achieved minimum margin and this bound. To illustrate this in the second experiment, we set L at 150 and tune P to 3/5, 2/3, 3/4, 4/5 and 9/10. In Figs. 2.1–2.6, we plot the relationship between each measure and the minimum margin with different values of P in one figure (Bigger versions of these figures can be found in our online supplementary material[2]). The figures show that increasing average accuracy does increase the upper-bound of the minimum margin, but the realized minimum margin may or may not increase (although the positive correlation between diversity and minimum margin is obvious in Fig. 2). In order to maximize the minimum margin of an ensemble, we need to maximize P and simultaneously satisfy the uniformity condition in Eq. (21). However, we have presented that the diversity measures cannot represent the uniformity condition well. Hence, we cannot expect a monotonic relationship between the generalization performance and the interactions between accuracy and diversity. This explains why experimental results in the past were inconclusive and contradictory. It was usually deemed that there exists a trade-off between the average accuracy and diversity. In other words, smaller P may correspond to larger *div*. This hypothesis is not reasonable according to our decomposition of the diversity measures. From Eqs. (15)–(20), all the diversity measures described in Section 2.1 can be formulated as $div = a - (bP + c\sum_{i=1}^{N} l_i^2)$, where a, b and c are constants. Hence relationship between *div* and P is influenced by the term $\sum_{i=1}^{N} l_i^2$. To study the relationship between P and $\sum_{i=1}^{N} l_i^2$, we carried out the third set of experiments. We set L at 15, randomly choose values for P between [0.5, 1] and generate ensembles with the chosen P. 10000 ensembles are generated again and relationship between P and the term $\sum_{i=1}^{N} l_i^2$ is plotted in Fig. 3. An obvious negative correlation between P and the term $\sum_{i=1}^{N} l_i^2$ can be observed with a correlation coefficient value of –0.9689. The strong negative correlation between P and $\sum_{i=1}^{N} l_i^2$ implies that a smaller P is not likely to result in a smaller value of $(bP + c\sum_{i=1}^{N} l_i^2)$, and hence may not correspond to larger *div*. Since the term $\sum_{i=1}^{N} l_i^2$ can take many different values for a given P, the true relationship between diversity and accuracy is complex. The answer to the second question presented in Section 1 is that we do not have to sacrifice some accuracy for diversity.

Another property we have observed about the existing diversity measures is the lack of a regularization term. Regularization is one of the key issues to be considered when designing a classification system. Since we never know true distribution of the data in real-world problems, the over-fitting problem of a classifier or classifier ensemble is naturally a major concern. In particular, the boosting type methods and SVM, two main algorithms that are

[2] The supplementary materials are available at: http://www.ntu.edu.sg/home/epnsugan/index_files/papers/JMLsup.pdf

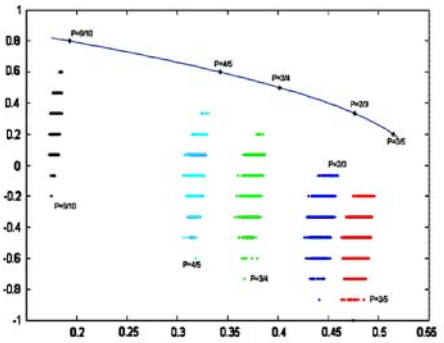

Figure 2.1: Disagreement measure

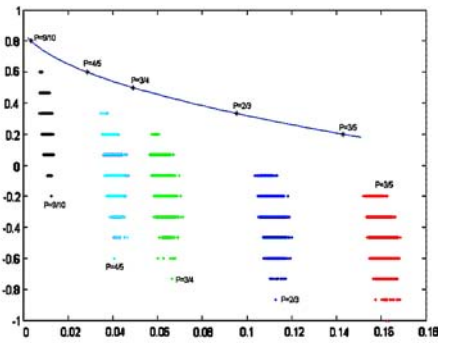

Figure 2.2: Double-fault measure

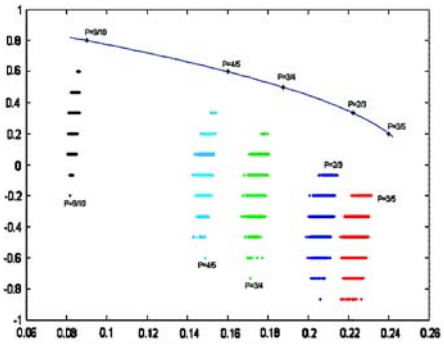

Figure 2.3: Kahavi-Wolpert variance

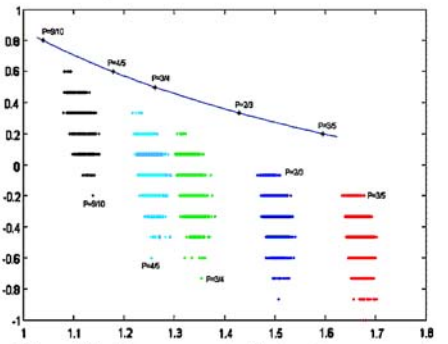

Figure 2.4: Measurement of interrater agreement

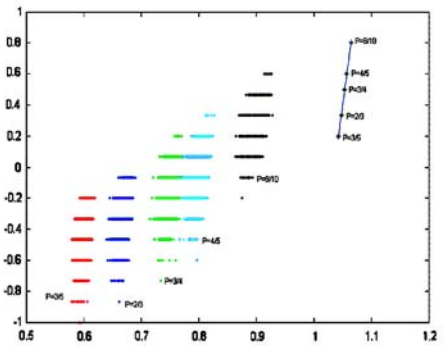

Figure 2.5: Generalized diversity

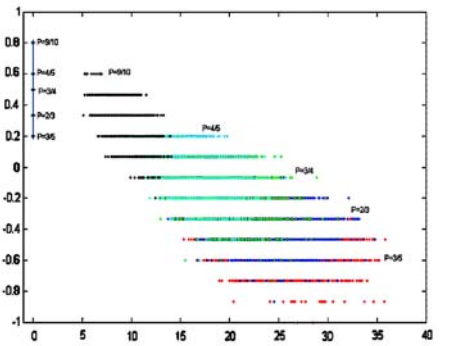

Figure 2.6: Measure of "Difficulty"

Fig. 2 The figures show the relationships between diversity measures and the minimum margin. $L = 15$, horizontal axis represents the diversity measures and vertical axis represents the minimum margin. The upper bound of the minimum margin is also shown in the figures corresponding to $P = 3/5$, $2/3$, $3/4$, $4/5$, and $9/10$ with "*"

based on the concept of margin maximization, may also be over-trained on noisy training data. Hence, regularization terms are incorporated in both boosting and SVM algorithms (Vapnik, 1995; Burges, 1998; Ratsch et al., 2001). In contrast, we cannot find a regularization term in the discussed six diversity measures. This observation implies that even when maximum values for the diversity measures are achieved, we may only obtain undesirably over-fitted solutions.

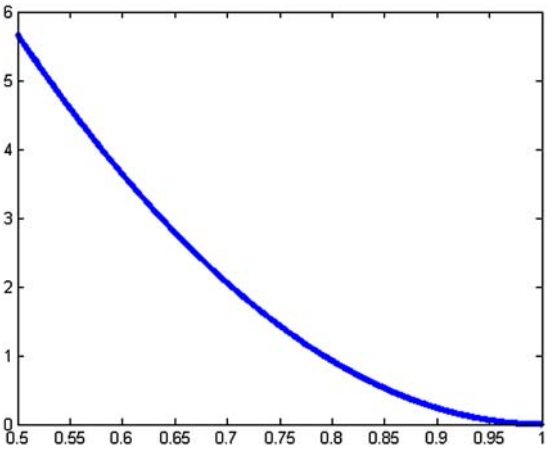

Fig. 3 Relationship between P and $\sum_{i=1}^{N} l_i^2$. Horizontal axis represents P and vertical axis is $\sum_{i=1}^{N} l_i^2$

All the experiments above do not really construct classifier ensembles and evaluate their performance on real-world problems. Hence we also carried out a simple empirical experiment to demonstrate our analysis. In this experiment, we employ least squares support vector machine (LSSVM) (Suykens et al., 2002) as the base classifier, and the ensembles are evaluated on 11 different 2-class datasets described in the work of Rätsch et al. (2001). For each of the datasets, the experiment was conducted for 100 times on randomly partitioned training and test sets with 60% for training and 40% for testing. For each pair of training and test sets, 300 different LSSVMs are generated. We first select the classifier with the best performance. Then we employ the diversity measures to sequentially select other base classifiers. Pseudo code of the experiment is presented below:

1. Generate 300 base classifiers using the training data.
2. Select the base classifier h_1 which has the smallest training error, the oracle output of this classifier is denoted by $\mathbf{O}_1$, $\mathbf{O} = \mathbf{O}_1$
3. Greedily select other base classifiers
   For $i = 1{:}149$
   for $j =$ all the unselected classifier
   calculate $div([\mathbf{O}, \mathbf{O}_j])$;
   select classifier h_j such that $div([\mathbf{O}, \mathbf{O}_j])$ is maximized;
   end
   $\mathbf{O} = [\mathbf{O}, \mathbf{O}_j]$,
   End
4. Apply the ensemble to test data using majority vote rule.

Results of the experiment are presented in Table 2. Among all the 11 datasets, exploiting diversity to construct ensemble only achieved better performance on breast cancer and thyroid datasets. For the other nine datasets, sometimes results of the ensembles are comparable to that of single LSSVM and SVM (e.g. *GD* for heart dataset, *DF*, *k* and *diff* for twonorm dataset). But we can observe that the ensembles perform significantly worse in many cases, such as *Dis* for heart dataset and *k* for diabetis dataset. According to previous analysis, these empirical results show that seeking diversity explicitly is not likely to generate promising classification performance consistently.

Table 2 Classification results of several well-known classification methods and ensembles achieved by seeking diversity explicitly. SVM is support vector machine and LSSVM is least-squares support vector machine*

	Dis	DF	KW	k	GD	diff	LSSVM	SVM
Banana	15.3 ± 1.4	12.7 ± 0.9	15.3 ± 1.4	13.4 ± 0.7	12.6 ± 0.7	13.1 ± 1.2	$\mathbf{10.8} \pm 0.6$	11.6 ± 0.7
Bcancer	25.8 ± 4.6	25.4 ± 4.0	25.8 ± 4.6	25.4 ± 4.8	25.4 ± 4.0	$\mathbf{25.3} \pm 4.5$	26.8 ± 4.5	26.0 ± 4.7
Diabetis	28.7 ± 2.0	24.7 ± 1.7	28.7 ± 2.0	28.9 ± 2.1	27.2 ± 1.7	26.9 ± 1.6	$\mathbf{23.5} \pm 1.8$	$\mathbf{23.5} \pm 1.7$
Fsolar	35.2 ± 1.6	35.1 ± 1.8	35.2 ± 1.9	35.2 ± 2.7	35.3 ± 1.7	35.2 ± 2.0	34.2 ± 1.9	$\mathbf{32.4} \pm 1.8$
German	25.1 ± 2.2	25.3 ± 2.2	25.1 ± 2.2	26.1 ± 2.1	25.9 ± 2.0	25.8 ± 2.0	$\mathbf{23.6} \pm 2.1$	$\mathbf{23.6} \pm 2.1$
Heart	26.2 ± 4.7	16.5 ± 3.6	26.2 ± 4.5	20.2 ± 3.5	17.5 ± 2.9	18.2 ± 3.2	16.4 ± 3.1	$\mathbf{16.0} \pm 3.3$
Ringnorm	6.5 ± 0.7	4.3 ± 0.8	6.5 ± 0.7	3.4 ± 0.3	3.5 ± 0.3	3.3 ± 0.2	$\mathbf{1.6} \pm 0.2$	1.7 ± 0.1
Thyroid	7.4 ± 3.7	4.9 ± 2.3	7.4 ± 3.3	4.9 ± 2.3	$\mathbf{4.5} \pm 2.0$	4.9 ± 2.2	4.9 ± 2.0	4.8 ± 2.2
Titanic	24.8 ± 4.7	22.7 ± 1.2	24.8 ± 0.7	23.2 ± 0.8	22.7 ± 1.2	24.7 ± 1.7	22.6 ± 0.8	$\mathbf{22.4} \pm 1.0$
Twonorm	8.4 ± 3.5	3.2 ± 0.3	8.4 ± 3.8	3.1 ± 0.2	3.5 ± 0.3	3.1 ± 0.3	$\mathbf{2.5} \pm 0.2$	3.0 ± 0.2
Waveform	18.5 ± 1.3	11.8 ± 0.7	18.5 ± 1.5	13.0 ± 0.7	12.3 ± 0.8	12.2 ± 0.6	$\mathbf{9.9} \pm 0.5$	$\mathbf{9.9} \pm 0.4$

*The results of regularized ADABOOSTING and SVM are originally presented by Rätsch et al. (2001).

Based on all the results presented above, we are able to make the following conclusions: Compared to those algorithms that seek diversity implicitly, exploiting diversity measures to seek diversity explicitly is ineffective in consistently achieving ensembles with good generalization performance. Firstly, the change of measured diversity cannot provide consistent guidance on whether a set of base classifiers possesses good generalization performance. Secondly, the measures are naturally correlated to the average accuracy of the base classifiers. This property is not desirable since we do not require the diversity measures to become another estimate for the classification accuracy. Finally, all the discussed diversity measures contain no regularization term. Therefore, even if these existing diversity measures can be maximized, the achieved ensemble may be an over-fit. It should be noted that the diversity measures are still useful. A more detailed discussion on exploiting the diversity measures is presented in Section 4.

3.3. Relationship between diversity and margins of an ensemble (General case)

So far we have presented our analysis based on the assumption that all the individual base classifiers are assigned the same weights. Now we will show that this assumption is not strictly necessary. Lemmas 1 and 2 can be generalized to the non-uniform weights of the base classifiers, which also validates our experimental analysis in Section 3.2 for the non-uniformly weighted case. For brevity we present the disagreement measure as an example here, detailed derivation for the other five measures can be found in Appendix B.

Since the weights of base classifiers are no longer same now, some equations we introduced in previous sections need to be modified. Given L base classifiers, the N-by-L oracle output matrix $\mathbf{O}$ and a weighting vector $\mathbf{w} = [w_1, w_2, \ldots, w_L]^T$ (where $w_j \geq 0$ and $\sum w_j = 1$) for the base classifiers, Eq. (5) is modified as:

$$dis = \frac{2}{L(L-1)} \sum_{j=1}^{L} \sum_{k=j+1}^{L} w_i w_j dis_{j,k} \tag{27}$$

and definition of l_i remains as:

$$l_i = L \sum_{O_{ij}=-1} w_j \tag{28}$$

With the definition presented by Eqs. (1) and (14), Eqs. (2) and (23) still hold in this case. Using some simple derivations, we have:

$$\sum_{j=1}^{L} \sum_{k=j+1}^{L} w_i w_j (n_{j,k}(1,-1) + n_{j,k}(-1,1)) = \frac{N\mathbf{w}^T\mathbf{w} - \mathbf{w}^T\mathbf{O}^T\mathbf{O}\mathbf{w}}{4} \tag{29}$$

Let $\{\mathbf{Ow}\}_i$ be the ith element of vector $\mathbf{Ow}$. For each training sample $\mathbf{x}_i$, we have

$$\{\mathbf{Ow}\}_i = \sum_{O_{ij}=1} w_j - \sum_{O_{ij}=-1} w_j$$

Then

$$l_i = L \sum_{O_{ij}=-1} w_j = \frac{L(1 - \{\mathbf{Ow}\}_i)}{2}$$

and

$$L - l_i = L - L \sum_{O_{ij}=-1} w_j = \frac{L(1 + \{\mathbf{Ow}\}_i)}{2}$$

Which gives

$$\sum_{i=1}^{N} l_i(L - l_i) = \frac{L^2(N - \mathbf{w}^T\mathbf{O}^T\mathbf{O}\mathbf{w})}{4} \tag{30}$$

So Eq. (27) can be rewritten as:

$$dis = \frac{2}{NL^3(L-1)} \sum_{i=1}^{N} l_i(L - l_i) - \frac{1 - \mathbf{w}^T\mathbf{w}}{2L(L-1)}$$

Since Eq. (2) still holds in this case, we have

$$\sum_{i=1}^{N} l_i = NL(1 - P)$$

and hence

$$dis = \frac{2(1-P)}{L(L-1)} - \frac{2}{NL^3(L-1)} \sum_{i=1}^{N} l_i^2 - \frac{1 - \mathbf{w}^T\mathbf{w}}{2L(L-1)} \tag{31}$$

As $\mathbf{w}$ is the given weighting vector, the term $\frac{1-\mathbf{w}^T\mathbf{w}}{2L(L-1)}$ can be regarded as a constant when optimizing *dis* with respect to l_i. Therefore, except for an additional L^2 term in the denominator, the equation above is almost the same as Eq. (15). When optimizing Eq. (31), since both the cost function and the constraint are in the same form as we have shown for majority vote case in Appendix A, the solution of this problem will also in the same form as in the majority case. Therefore, the previous analysis is also applicable to non-uniformly weighted case, and Lemma 2 can be generalized as:

Lemma 3 (generalized form of Lemma 2). *Given L base classifiers weighted by weights* w_1, $w_2, \ldots, w_L$, *if P is regarded as a constant and the maximum diversity is achievable, maximizing diversity among the base classifiers is equivalent to maximizing the minimum margin of the ensemble on the training samples.*

4. Discussions

Up to this point, we have analyzed six diversity measures. All of them require satisfying the uniformity condition. When the uniformity condition is maximized, the minimum margin of the classifier ensemble is maximized. In order to design a classifier ensemble, simply analyzing relationship between diversity, training accuracy and margin of the ensemble is not enough. All the conclusions of the theoretical and experimental results must be considered in a much larger context. Hence, in this section we try to propose some answers for the third question presented in Section 1 by examining possible applications of diversity measures in an ensemble learning algorithm.
A typical ensemble learning algorithm can be summarized in three stages:

1. Given the training samples, generate a set of base classifiers
2. Select a subset of the generated base classifiers
3. Construct the ensemble with a combination scheme

These procedures can be illustrated using two examples, a boosting algorithm and the neural network ensemble proposed by Giacinto and Roli (2001). In a boosting algorithm, we first generate the base classifiers sequentially. After that, the base classifiers are pruned either to avoid over-fitting or to reduce computational complexity for testing. Here pruning functions as a classifier selector and several methods have been proposed in the literature (Margineantu & Dietterich, 1997; Tamon & Xiang, 2000). Finally, the selected base classifiers are combined according to the predefined weighted vote rule. In the second example, after generating the base classifiers, a clustering method is employed to cluster them, then only one classifier is chosen from each cluster to construct the ensemble, and majority vote is employed as the combination rule.

In these procedures, a classifier ensemble is actually constructed in stages 2 and 3. What is the optimal base classifier is unclear in stage 1 and any base classifier generated in this stage may be useful for the final ensemble. Therefore, the goal in this stage is only to generate diverse classifiers rather than achieving a good classifier ensemble. Bagging, boosting and random subspace methods can all be employed to seek diversity in this stage. Bagging and random subspace methods generate different classifiers by varying the training data randomly, while boosting achieves diversity by using a deterministic procedure. In this situation, one can also employ diversity measures by including them in the objective function. Since varying training data (implemented as bagging or random subspace method) may not influence performance of the classifier substantially, exploiting a diversity measure is likely to generate more diverse

classifiers because difference between classifiers is required more explicitly. Whether the definition of diversity is precise or not is not very important in this stage.

The problem to be solved in the second stage is much more difficult than in the first stage. If one exploits diversity measures as criteria to select the base classifiers, then the diversity measure is required to be precise, since the choice of diversity measure will directly influence the final ensemble and subsequently the classification result. Unfortunately, as shown in the previous section, none of the six diversity measures is suitable for this task. In addition to all the theoretical analyses already presented, another problem that is important for practical implementation needs to be noted.

Matrix Cover Problem

Input: A positive constant θ and a real-valued matrix $\mathbf{O}$ of size N-by-T such that, for any integer $1 \leq i \leq N$, $\sum_{j=1}^{T} O_{ij} > \theta$

Output: A minimal subset L of the columns of $\mathbf{O}$ such that, for all $1 \leq i \leq N$, $\sum_{j=1}^{L} O_{ij} > \theta$

By replacing θ with the minimum margin calculated from L and P, the base classifier selection problem can be formulated as a matrix cover problem. Tamon and Xiang (2000) proved that the Matrix Cover Problem below is NP-complete. Hence, we can hardly achieve the optimal subset of base classifiers corresponding to the optimal margin. As the diversity measures are not single-valued functions except for the optimized case (i.e. Eq. (21) is satisfied), different ensembles may have the same P and diversity *div* but different performance on the data. One cannot differentiate between these kinds of ensembles by employing diversity measures, which will cause problems in the implementation of the algorithm. On the other hand, the base classifier selection problem has been well studied in the pattern recognition literature as a feature selection problem. Hence, instead of using diversity measures, we can employ a well-developed feature selection method to select base classifiers, such as sequential forward (as we employed in the experiment on real-world datasets) and sequential floating forward scheme with different selection criteria.

After selecting L base classifiers, a combination scheme should be decided in the third stage. This problem is naturally a classification problem in an L-dimensional space. Hence effective classification methods can be easily exploited. Diversity measures are generally not applicable in this stage. Another application of diversity measures is to visualize relationships of the base classifiers in an ensemble or different ensembles. This application is not directly related to ensemble design. Hence, the precise definition of diversity is also not quite important.

A frequently asked question is: What is a "good" diversity measure for designing an ensemble learning algorithm? We have discussed three possible applications of the diversity measures in an ensemble learning algorithm: generating individual classifiers, visualizing relevant properties of the ensemble and selecting base classifiers. In our opinion, the existing diversity measures are sufficient for the first two purposes, but are not for the last one. Our analysis partially suggests some evaluation criteria for proposing new diversity measures. As maximizing the minimum margin of the ensemble (or satisfying the uniformity condition) is the objective of maximizing diversity, one will expect the minimum margin to be a single-valued function of diversity and expect it to asymptotically converge to the maximum value. Different objectives other than the margin can also be used, but no matter what it is, it should be monotonic and single-valued with respect to the measure.

5. Conclusions

In the pattern recognition field, many good algorithms have heuristically been proposed. Subsequently, theoretical analyses are developed to explain the good performance and further improve them. Reviewing the literature of ensemble learning, one can find several theories that tend to explain the success of ensemble learning algorithms, such as the concept of diversity, the concept of margin, the ambiguity decomposition (Krogh & Vedelsby, 1995), the bias-variance decomposition and the bias-variance-covariance decomposition (Liu & Yao, 1997). According to the numerous previous works, all of these theories do present some of the underlying mechanism of ensemble learning, and thus should be related to one another. However, to our knowledge, such studies mainly exist in the regression context (Brown et al., 2004).

We have reviewed six existing measures that quantify the diversity among base classifiers of a classifier ensemble, and demonstrated explicitly the relation between these measures and the margin maximization concept, which accounts for the success of several pattern classification algorithms. Since the diversity measures are motivated from different areas of pattern classification, identifying the link between these measures is required to study why these measures are useful or not for classification. In this process, we presented the uniformity condition for maximizing both the diversity and the minimum margin of an ensemble and demonstrated theoretically and experimentally the ineffectiveness of the diversity measures for constructing ensembles with good generalization performance. We believe that our analysis has highlighted some relationships between different theories in the classification context, and could hence help design better ensemble learning algorithms.

In addition to the presented six diversity measures, many existing diversity measures have originated from different research areas that remotely relate to pattern recognition field rather than from pattern recognition field itself. It is natural that not all of them can be put into one framework,[3] and a thorough analysis of all the diversity measures is out of the scope of this work. However, previous experimental studies have shown that most diversity measures perform similarly. Thus we can expect other diversity measures to have similar properties as the measures analyzed in this work.

Appendix A

Proofs of Eqs. (15)–(20) *and Lemma 1*

Proof for the disagreement measure:

As defined in Section 2, for a sample $\mathbf{x}_i$, l_i base classifiers classify it incorrectly and the other $L- l_i$ classifiers classify it correctly. Then for this sample, there are $l_i(L- l_i)$ pairs of base classifiers whose oracle outputs are different. Hence, the term $\sum_{j=1}^{L}\sum_{k=j+1}^{L}(n_{j,k}(1,-1)+n_{j,k}(-1,1))$ is equivalent to the term$\sum_{i=1}^{N} l_i\,(L-l_i)$, and Eq. (6) can be re-written as:

$$dis = \frac{2}{NL(L-1)}\sum_{i=1}^{N} l_i\,(L-l_i)$$

$$= \frac{2}{N\,(L-1)}\sum_{i=1}^{N} l_i - \frac{2}{NL(L-1)}\sum_{i=1}^{N} l_i^2. \tag{A1}$$

[3] In the online supplementary materials, we also discuss two more diversity measures that are relevant to our analysis.

We derive from Eq. (2) that:

$$\sum_{i=1}^{N} l_i = NL(1-P). \tag{A2}$$

To maximize the diversity, *dis* need to be maximized. By substituting Eq. (A2) into Eq. (A1), we get:

$$dis = \frac{2L(1-P)}{(L-1)} - \frac{2}{NL(L-1)} \sum_{i=1}^{N} l_i^2. \tag{15}$$

If P is regarded as a constant, the diversity maximization problem becomes a Lagrangian formulation: We are given the constrained maximization problem:

Maximize

$$dis = \frac{2L(1-P)}{(L-1)} - \frac{2}{NL(L-1)} \sum_{i=1}^{N} l_i^2$$

With respect to the constraint

$$\sum_{i=1}^{N} l_i = NL(1-P).$$

By introducing the Lagrangian multiplier, we obtain

$$Ldis = \frac{2L(1-P)}{(L-1)} - \frac{2}{NL(L-1)} \sum_{i=1}^{N} l_i^2 + \lambda \cdot \sum_{i=1}^{N} l_i - \lambda \cdot NL(1-P). \tag{A3}$$

For all i, differentiating Eq. (A3) with respect to l_i, we obtain

$$\frac{d(Ldis)}{d(l_i)} = -\frac{4l_i}{NL(L-1)} + \lambda = 0. \tag{A4}$$

Hence, for all i:

$$l_i = \frac{\lambda NL(L-1)}{4}, \tag{A5}$$

which means

$$\lambda = \frac{1-P}{N(L-1)} \tag{A6}$$

and

$$l_i = L(1-P). \tag{21}$$

Therefore, Eq. (15) is maximized when Eq. (21) is satisfied.

$$\max(dis) = \frac{2LP(1-P)}{L-1} \tag{A7}$$

if and only if:

$$l_i = L(1-P)\,\forall i. \tag{21}$$

Proof for the double-fault measure:
According to the definition, for each sample $\mathbf{x}_i$, there are $l_i(l_i-1)/2$ pairs of base classifiers whose oracle outputs on x_i are -1. By observing that the term $\sum_{j=1}^{L}\sum_{k=j+1}^{L} n_{j,k}(-1,-1)$ is equivalent to the term $\sum_{i=1}^{N} l_i(l_i-1)/2$, we re-write Eq. (8) as:

$$\begin{aligned} DF &= \frac{1}{NL(L-1)}\sum_{i=1}^{N} l_i(l_i-1) \\ &= \frac{1}{NL(L-1)}\sum_{i=1}^{N} l_i^2 - \frac{1-P}{L-1}. \end{aligned} \tag{16}$$

In this case, we need to minimize the *DF* to maximize the diversity. Similar to the proof for *dis*, after solving the minimization problem by introducing a Lagrangian multiplier, we obtain:

$$\min(DF) = \frac{(1-P)(L-LP-1)}{L-1} \tag{A8}$$

if and only if Eq. (21) is satisfied.

Proof for the Kohavi-Wolpert variance
From the proof for the first two measures, it is easy to derive that:

$$\begin{aligned} KW &= \frac{1}{NL^2}\sum_{i=1}^{N} l_i(L-l_i) \\ &= 1 - P - \frac{1}{NL^2}\sum_{i=1}^{N} l_i^2. \end{aligned} \tag{17}$$

We need to maximize Eq. (17) to maximize the diversity. Then the proof is exactly the same as the proof for the disagreement measure. If and only if Eq. (21) is satisfied, the maximum diversity can be achieved as:

$$\max(KW) = P(1-P). \tag{A9}$$

Proof for the measurement of inter-rater agreement:
It can be derived directly from the definition of this measure that:

$$k = 1 - \frac{\sum_{i=1}^{N}(L-l_i)\,l_i}{NL(L-1)\,P(1-P)}$$

$$= 1 - \frac{NL^2(1-P) - \sum_{i=1}^{N} l_i^2}{NL(L-1)P(1-P)}$$

$$= \frac{LP - P - L}{LP - P} + \frac{\sum_{i=1}^{N} l_i^2}{NL(L-1)P(1-P)}. \tag{18}$$

The maximum diversity can be achieved when Eq. (18) is minimized, which means $\sum_{i=1}^{N} l_i^2$needs to be minimized. Similar to the former three proofs, we only need to satisfy Eq. (21) and:

$$\min(k) = -\frac{1}{L-1}. \tag{A10}$$

Proof for the generalized diversity:
We first prove that the terms $\sum_{j=1}^{L} \frac{j}{L} T_J$ and $\sum_{j=1}^{L} \frac{j(j-1)}{L(L-1)} T_j$ are equivalent to the terms $1 - P$ and $\frac{\sum_{i=1}^{N} l_i(l_i-1)}{NL(L-1)}$ respectively:
Let *n(j)* be the number of samples classified incorrectly by *j* base classifiers, then:

$$\sum_{j=0}^{L} n(j) = N \tag{A11}$$

$$T_j = \frac{n(j)}{N}. \tag{A12}$$

From the definition of l_i, we can get

$$\sum_{j=1}^{L} \frac{j}{L} T_j = \sum_{j=1}^{L} \frac{j}{L} T_j + \frac{0}{L} T_0$$

$$= \sum_{j=0}^{L} \frac{n(j)\, j}{NL}$$

$$= \frac{1}{NL} \sum_{j=0}^{L} \sum_{i=1, l_i=j}^{N} l_i. \tag{A13}$$

Since

$$\sum_{j=0}^{L} \sum_{i=1, l_i=j}^{N} l_i = \sum_{i=1}^{N} l_i, \tag{A14}$$

We can obtain:

$$\sum_{j=1}^{L} \frac{j}{L} T_j = \frac{\sum_{i=1}^{N} l_i}{NL} = 1 - P \tag{A15}$$

Similarly,

$$\begin{aligned}\sum_{j=0}^{L} &= \frac{j\,(j-1)}{L\,(L-1)}T_j = \sum_{j=0}^{L}\frac{j\,(j-1)}{L\,(L-1)}T_j \\ &= \frac{1}{NL\,(L-1)}\sum_{j=0}^{L} j\,(j-1)\,n\,(j) \\ &= \frac{1}{NL\,(L-1)}\sum_{j=0}^{L}\sum_{i=1,l_i=j}^{n} l_i\,(l_i-1) \\ &= \frac{\sum_{i=1}^{N} l_i\,(l_i-1)}{NL\,(L-1)}. \end{aligned} \tag{A16}$$

From Eqs. (A15) and (A16), Eq. (12) can be re-written as:

$$\begin{aligned} GD &= 1 - \frac{\sum_{j=1}^{L}\frac{j(j-1)}{L(L-1)}T_j}{\sum_{j=1}^{L}\frac{j}{L}T_j} \\ &= 1 - \frac{\sum_{i=1}^{N} l_i\,(l_i-1)}{NL\,(L-1)\,(1-P)} \\ &= \frac{L}{L-1} - \frac{\sum_{i=1}^{N} l_i^2}{NL\,(L-1)\,(1-P)}. \end{aligned} \tag{19}$$

If and only if Eq. (21) is satisfied, the generalized diversity is maximized:

$$\max\,(GD) = \frac{LP}{L-1}. \tag{A17}$$

Proof for the measure of "Difficulty":
From the definition of this measure, we need to minimize the term:

$$\begin{aligned} V &= var\left(\frac{L-l_i}{L}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{NL}\sum_{j=1}^{N} l_j - l_i\right)^2 \\ &= \frac{1}{NL}\sum_{i=1}^{N} l_i^2 - L\,(1-P)^2\,. \end{aligned} \tag{20}$$

It is obvious that when Eq. (21) is satisfied

$$\min\,(V) = 0. \tag{A18}$$

Appendix B

In this section we show that our analysis can also be generalized to non-uniformly weighted case for the other diversity measures besides the disagreement measure.

Proof for Double-Fault measure:
Similar to Eq. (27), Eq. (8) is redefined as:

$$DF = \frac{2}{NL(L-1)} \sum_{j=1}^{L} \sum_{k=j+1}^{L} w_k w_j n_{j,k}(-1,-1) \tag{A19}$$

Let **1** be an N-by-L matrix with all the elements as one. It is easy to get:

$$\sum_{j=1}^{L} \sum_{k=j+1}^{L} w_k w_j n_{j,k}(-1,-1) = \frac{1}{2} \left(\frac{\mathbf{w}^T (\mathbf{O}-\mathbf{1})^T (\mathbf{O}-\mathbf{1}) \mathbf{w}}{4} - \sum_{j=1}^{L} w_j^2 N (1-p_j) \right)$$

Since there is:

$$\sum_{i=1}^{N} l_i^2 = \frac{L^2 \mathbf{w}^T (\mathbf{O}-\mathbf{1})^T (\mathbf{O}-\mathbf{1}) \mathbf{w}}{4} \tag{A20}$$

Eq. (A19) can be re-written as:

$$DF = \frac{1}{NL^3(L-1)} \left(\sum_{i=1}^{N} l_i^2 - \sum_{j=1}^{L} w_j^2 L^2 N (1-p_j) \right) \tag{A21}$$

Again, by optimizing DF with respect to l_i, we can validate our analysis in the non-uniformly weighted case. If w_j is one for all base classifiers, Eq. (A21) is almost the same as Eq. (16) except for an additional L^2 term in the denominator. Therefore, Eq. (16) is a special case of (A21).

Proof for the Kohavi-Wolpert variance
Following the same rationale described in Section 2.1.3, we have

$$P(O=-1 \mid \mathbf{x}) = \sum_{O_{ij}=-1} w_j = \frac{l_i}{L} \quad \text{and} \quad P(O=1 \mid \mathbf{x}) = \sum_{O_{ij}=1} w_j = 1 - \frac{l_i}{L}$$

Then we get

$$\begin{aligned} KW &= \frac{1}{2N} \sum_{i=1}^{N} \left(1 - \sum_{k=1}^{C} P(y=\omega_k \mid \mathbf{x})^2 \right) \\ &= \frac{1}{2N} \sum_{i=1}^{N} \left(1 - \frac{l_i^2}{L^2} - \left(1 - \frac{l_i}{L} \right)^2 \right) = \frac{1}{NL^2} \sum_{i=1}^{N} l_i (L - l_i) \end{aligned} \tag{A22}$$

(A22) is exactly same as Eq. (10), hence our analysis is equivalent for both uniformly and non-uniformly weighted case.

Proof for the measurement of inter-rater agreement and "Difficulty":
According to Eqs. (11) and (20), the definitions of these two measures directly include the term l_i and do not contain any additional constraints on $\mathbf{w}$. Therefore, Eqs. (11) and (20) are valid for both uniformly and non-uniformly weighted cases. Since Eq. (2) also remains unchanged in non-uniformly weighted case, it is obvious that any analysis based on l_i's in uniformly weighted case can be generalized to non-uniformly weighted case for these two measures, and the Lemma 3 holds.

Proof for the generalized diversity:
Because Eqs. (A15) and (A16) directly result in Eq. (19), to generalize Eq. (19) to non-uniformly weighted case, we only need to check whether Eqs. (A15) and (A16) still hold if the weights are non-uniform. If the weights are non-uniform, l_i and j in the original definition will no longer only take integers from 0 to L, but they are still discontinuous. We define Ω as the set that contains all possible values of l_i. Then Eq. (A11) can be re-written as (A23) and Eq. (A12) still holds.

$$\sum_{j\in\Omega} n(j) = N \tag{A23}$$

Further, Eqs. (A13) and (A14) can be modified as:

$$\sum_{j\in\Omega} \frac{j}{L} T_j = \frac{1}{NL} \sum_{j\in\Omega} \sum_{i=1, l_i=j}^{N} l_i \tag{A24}$$

$$\frac{1}{NL} \sum_{j\in\Omega} \sum_{i=1, l_i=j}^{N} l_i = \sum_{i=1}^{N} l_i \tag{A25}$$

Combining Eqs. (A24) and (A25) with Eq. (2), we can prove that (A15) still holds when the weights are non-uniform. Eq. (A16) can be shown to be true in a similar manner. Therefore, Eq. (19) also holds in non-uniformly weighted case and Lemma 3 is true for the generalized diversity measure.

Acknowledgments We thank the referees for their valuable comments.

References

Atukorale, A. S., Downs, T., & Suganthan, P. N. (2003). Boosting HONG Networks. *Neurocomputing*, *51*, 75–86.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning, 36*, 105-142.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45,* 5–32.

Brown, G., Wyatt, J. L., Harris, R. & Yao, X. (2004). Diversity Creation Methods: A Survey and Categorization. *Information Fusion Journal (Special issue on Diversity in Multiple Classifier Systems)*, *6*(1), 5–20.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning, 40*(2), 1–22.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. John Wiley & Sons.
Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, *121*(2), 256–285.
Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In: *Proc. 13th Int. Conference on Machine Learning* (pp. 148–156). Morgan Kaufmann.
Giacinto, G., & Roli, F. (2001). Design of effective neural network ensembles for image classification processes. *Image Vision and Computing*, *19*(9/10), 699–707.
Hansen, L. & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001.
Ho, T. (1998). The random space method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844.
Kohavi, R., & Wolpert, D. (1996). Bias plus variance decomposition for zero-one loss functions. In: L. Saitta (Ed.), *Proc. 13th Int. Conference on Machine Learning* (pp. 275–283). Morgan Kaufmann.
Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active Learning. In: G. Tesauro, D. S. Touretzky and T. K. Leen (Eds.), *Advances in Neural Information Processing Systems 7* (pp. 231–238). Cambridge, MA: MIT Press.
Kuncheva, L., & Whitaker, C. (2003a). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51,* 181–207.
Kuncheva, L.I. (2003b). That elusive diversity in classifier ensembles. In: *Proc IbPRIA 2003, Mallorca, Spain*, 2003, *Lecture Notes in Computer Science*, *Springer-Verlag*, LNCS 2652, 1126–1138.
Liu, Y., & Yao, X. (1997). Negatively correlated neural networks can produce best ensembles. *Australian Journal of Intelligent Information Processing Systems, 4*(3/4), 176–185.
Liu, Y., Yao, X., & Higuchi, T. (2000). Evolutionary ensembles with negative correlation learning. *IEEE Transactions on Evolutionary Computation, 4,* 380–387.
Margineantu, D., & Dietterich, T. (1997). Pruning Adaptive Boosting. In: *Proceedings ICML'97: International Conference on Machine Learning* (pp. 211–218). Los Altos, CA: Morgan Kaufmann.
Mason, L., Bartlett, P. L., & Baxter, J. (2000). Improved generalization through explicit optimization of margins. *Machine Learning*, *38*(3), 243–255.
Patridge, D., & Krzanowski, W. J. (1997). Software diversity: Practical statistics for its measurement and exploitation. *Information & Software Technology, 39*, 707–717.
Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for AdaBoost. *Machine Learning*, *42*(3), 287–320.
Schapire, R. E., Freund, Y., Bartlett, P. L., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, *26*(5), 1651–1686.
Schapire, R. (1999). Theoretical views of boosting. In: *Proc. 4th European Conference on Computational Learning Theory* (pp. 1–10).
Skalak, D. (1996). The sources of increased accuracy for two proposed boosting algorithms. In: *Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*.
Suganthan, P. N. (1999). Hierarchical Overlapped SOM's for Pattern Classification. *IEEE Transactions on Neural Networks*, *10*(1), 193–196.
Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least Squares Support Vector Machines Singapore*. World Scientific.
Tamon, C., & Xiang, J. (2000). On the Boosting Pruning problem. In: R. L. Mantaras and E. Plaza (Eds.), *Machine Learning: Proc. 11th European Conference*, Vol. 1810 Lecture Notes in Computer Science (pp. 404–412). Springer.
Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer.