Multi-Information Ensemble Diversity

Zhi-Hua Zhou[1] and Nan Li[2]

[1] National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
[2] School of Mathematical Sciences
Soochow University, Suzhou 215006, China
zhouzh@lamda.nju.edu.cn linan@suda.edu.cn

Abstract. Understanding ensemble diversity is one of the most important fundamental issues in ensemble learning. Inspired by a recent work trying to explain ensemble diversity from the information theoretic perspective, in this paper we study the ensemble diversity from the view of *multi-information*. We show that from this view, the ensemble diversity can be decomposed over the component classifiers constituting the ensemble. Based on this formulation, an approximation is given for estimating the diversity in practice. Experimental results show that our formulation and approximation are promising.

1 Introduction

It is well-known that in order to build a good ensemble, the component classifiers should be *accurate* as well as *diverse*. There are effective processes for estimating the accuracy of component classifiers, however, measuring diversity is not easy since there is no generally accepted formal definition. During the past decade, many diversity measures have been designed; to name a few, the *Q-statistics* [11], the *disagreement* [9], the *double-fault* [7], the κ-*statistic* [5], etc. However, it has been disclosed that existing diversity measures are suspect [10].

Recently, Brown [2] investigated the ensemble diversity from an information theoretic perspective. He found that the ensemble *mutual information* can be naturally expanded into 'accuracy' and 'diversity' terms, and the ensemble diversity exists at multiple orders of correlation. We believe that this is an important step towards the understanding of ensemble diversity. However, the expressions of that information theoretic diversity and its terms, especially the involved *interaction information*, are quite complicated, and there is no proposal of effective process for estimating the multiple orders of correlation in practice.

Inspired by Brown's work [2], in this paper we also study the ensemble diversity from an information theoretic perspective. From the view of *multi-information*, we propose a new formulation where the ensemble diversity and the related terms are simpler. This formulation enables to decompose the diversity over the component classifiers. Based on the formulation, we give an approximation for estimating the ensemble diversity in practice. Experiments show that our formulation and approximation are promising.

The rest of this paper is organized as follows. Section 2 briefly reviews some basics of information theory and Brown's study. Section 3 introduces our formulation based on multi-information. Section 4 presents an approximation. Section 5 reports on experiments. Finally, Section 6 concludes.

2 Background

The fundamental concept of information theory is the *entropy*, which is a measure of uncertainty. For a variable X, its entropy $H(X)$ is defined as $\sum_x p(x)\log(p(x))$, where x is the value of X, and $p(x)$ is the probability distribution.

Based on the concept entropy, the dependence among multiple variables can be measured by mutual information and its multivariate generalizations. Denote n variables $X_1, \cdots, X_n$ as $X_{1:n}$ and another variable as Y, then,

- *Mutual information* and *conditional mutual information* [4]:

$$I(X_1; X_2) = \sum\nolimits_{x_1,x_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \tag{1}$$

$$I(X_1; X_2 \mid Y) = \sum\nolimits_{y,x_1,x_2} p(y)p(x_1, x_2 \mid y) \log \frac{p(x_1, x_2 \mid y)}{p(x_1 \mid y)p(x_2 \mid y)} \tag{2}$$

- *Multi-information* and *conditional multi-information* [15, 14, 13]:

$$\mathcal{I}(X_{1:n}) = \sum\nolimits_{x_{1:n}} p(x_1, \cdots, x_n) \log \frac{p(x_1, \cdots, x_n)}{p(x_1)p(x_2)\cdots p(x_n)} \tag{3}$$

$$\mathcal{I}(X_{1:n} \mid Y) = \sum\nolimits_{y,x_{1:n}} p(y)p(x_{1:n} \mid y) \log \frac{p(x_{1:n} \mid y)}{p(x_1 \mid y)\cdots p(x_n \mid y)} \tag{4}$$

- *Interaction information* [12]:

$$I(\{X_{1:n}\}) = \begin{cases} I(X_1, X_2) & \text{for } n = 2 \\ I(\{X_{1:n-1}\} \mid X_n) - I(\{X_{1:n-1}\}) & \text{for } n \geq 3 \end{cases} \tag{5}$$

where $p(x_{1:n})$ is the joint distribution of $X_{1:n}$, $p(x)$ and $p(y)$ are the marginal distributions, and $p(\cdot \mid \cdot)$'s are the conditional distributions. [1]

As described above, *mutual information* measures the mutual dependence of two variables, while both *multi-information* and *interaction information* are its multivariate generation which express the dependence among multiple variables. Like mutual information, multi-information is nonnegative and equals zero if and only if all the variables are independent. Interaction information, however, can be negative; this has likely encumbered its wide application as an information measure.

In ensemble learning, suppose there is a set of classifiers $S = \{X_1, \cdots, X_m\}$ and the target class is Y; our objective is to find a combination function g

[1] We will use $p(\cdot)$ to denote different distributions when the meaning is clear.

that minimizes the probability of error prediction $p\left(g\left(X_{1:m}\right) \neq Y\right)$. Brown [2] bounded the probability of error by two inequalities [6, 8], that is,

$$\frac{H(Y) - I(X_{1:m}; Y) - 1}{\log(|Y|)} \leq p\left(g\left(X_{1:m}\right) \neq Y\right) \leq \frac{H(Y) - I(X_{1:m}; Y)}{2}. \quad (6)$$

Thus, to minimize the prediction error, the mutual information $I(X_{1:m}; Y)$ should be maximized. Subsequently, an expansion of $I(X_{1:m}; Y)$ was given based on [3, Thm 1], and the *information theoretic diversity*, *redundancy* and *conditional redundancy* were defined. Denote T_k as a set of size k. Then,

$$\begin{aligned} I(X_{1:m}; Y) &= \underbrace{\sum\nolimits_{i=1}^{m} I(X_i; Y)}_{\text{relevancy}} + \underbrace{\sum\nolimits_{k=2}^{m} \sum\nolimits_{T_k \subseteq S} I(\{T_k \cup Y\})}_{\text{information theoretic diversity}} \quad (7) \\ &= \sum\nolimits_{i=1}^{m} I(X_i; Y) + \underbrace{\sum\nolimits_{k=2}^{m} \sum\nolimits_{T_k \subseteq S} I(\{T_k\}|Y)}_{\text{conditional redundancy}} - \underbrace{\sum\nolimits_{k=2}^{m} \sum\nolimits_{T_k \subseteq S} I(\{T_k\})}_{\text{redundancy}}. \end{aligned}$$

As shown above, the information theoretic diversity naturally emerges as an expression of the interaction information. It can also be found that the ensemble diversity exists at multiple orders of correlation. Since computing high-order interaction information is generally difficult, only pairwise interactions were monitored in [2]. Such a pairwise diversity, however, can only capture the low-order components of the multiple orders of correlation.

3 Multi-Information Diversity

Lemma 1. *The multi-information and conditional multi-information can be expanded as a sum of mutual information and conditional mutual information terms, respectively. That is,*

$$\mathcal{I}(X_{1:n}) = \sum\nolimits_{i=2}^{n} I(X_i; X_{1:i-1}); \quad (8)$$

$$\mathcal{I}(X_{1:n} \mid Y) = \sum\nolimits_{i=2}^{n} I(X_i; X_{1:i-1} \mid Y). \quad (9)$$

Proof. The multi-information can be written as [4, 13]

$$\begin{aligned} \mathcal{I}(X_{1:n}) &= \sum\nolimits_{i=1}^{n} H(X_i) - H(X_{1:n}) \\ &= \sum\nolimits_{i=1}^{n} H(X_i) - \left[H(X_1) + \sum\nolimits_{i=2}^{n} H(X_i \mid X_{1:i-1})\right] \\ &= \sum\nolimits_{i=2}^{n} \left(H(X_i) - H(X_i \mid X_{1:i-1})\right) = \sum\nolimits_{i=2}^{n} I(X_i; X_{1:i-1}) \ , \end{aligned}$$

which is the result in Eq. 8.

For conditional multi-information, its definition in Eq. 4 can be written as

$$\sum\nolimits_{y, x_{1:n}} p(x_{1:n}, y) \left[\log \frac{p(x_{1:n}, y)}{p(x_1) \cdots p(x_n) p(y)} - \log \frac{p(x_1, y) \cdots p(x_n, y)}{p(x_1) \cdots p(x_n) p(y)^n}\right].$$

Then, by Eqs. 1 and 3, we have

$$\begin{aligned}\mathcal{I}(X_{1:n} \mid Y) &= \mathcal{I}(X_{1:n}, Y) - \sum_{i=1}^{n} I(X_i; Y) \\ &= \sum_{i=2}^{n} [H(X_i, Y) + H(X_{1:i-1}, Y) - H(X_{1:i}, Y) - H(Y)] \\ &= \sum_{i=2}^{n} I(X_i; X_{1:i-1} \mid Y),\end{aligned}$$

which completes the proof. □

Theorem 1. *For a set of m classifiers $S = \{X_1, \cdots, X_m\}$ and the class label Y, the mutual information $I(X_{1:m}; Y)$ can be expanded as*

$$\begin{aligned}I(X_{1:m}; Y) &= \underbrace{\sum_{i=1}^{m} I(X_i; Y)}_{\text{relevance}} + \underbrace{\mathcal{I}(X_{1:m} \mid Y) - \mathcal{I}(X_{1:m})}_{\text{multi-information diversity}} \\ &= \sum_{i=1}^{m} I(X_i; Y) + \underbrace{\sum_{i=1}^{m} I(X_i; X_{1:i-1} \mid Y)}_{\text{conditional redundancy}} - \underbrace{\sum_{i=1}^{m} I(X_i; X_{1:i-1})}_{\text{redundancy}}.\end{aligned} \tag{10}$$

Proof. Based on the properties of mutual information, we have

$$\begin{aligned}I(X_{1:m}; Y) &= H(X_{1:m}) + H(Y) - H(X_{1:m}, Y) \\ &= \sum_{i=1}^{m} H(X_i) + H(Y) - H(X_{1:m}, Y) + H(X_{1:m}) - \sum_{i=1}^{m} H(X_i) \\ &= \mathcal{I}(X_{1:m}, Y) - \mathcal{I}(X_{1:M})\end{aligned}$$

By adding $\sum_{i=1}^{m} I(X_i; Y) - \sum_{i=1}^{m} I(X_i; Y)$ to the right-hand side of above equation, it follows that

$$\begin{aligned}I(X_{1:m}; Y) &= \sum_{i=1}^{m} I(X_i; Y) + \mathcal{I}(X_{1:m}, Y) - \sum_{i=1}^{m} I(X_i; Y) - \mathcal{I}(X_{1:m}) \\ &= \sum_{i=1}^{m} I(X_i; Y) + \sum_{i=1}^{m} I(X_i; X_{i-1}, \cdots, X_1 \mid Y) - \mathcal{I}(X_{1:m}).\end{aligned}$$

From Eq. 8 in Lemma 1, $\mathcal{I}(X_{1:m})$ can be written as $\sum_{i=1}^{m} I(X_i; X_{1:i-1})$. Therefore, above equation becomes Eq. 10, which completes the proof. □

Comparing with Eq. 7, the formulation of Eq. 10 is much simpler while the meanings are easier to understand; that is, the *redundancy* and *conditional redundancy* are multi-information and conditional multi-information, and the multi-information diversity is just their difference.

Both the *redundancy* and *conditional redundancy* in Eq. 10 are in the form of sum of $I(X_i; X_{1:i-1})$'s and $I(X_i; X_{1:i-1} \mid Y)$'s, respectively. One advantage of our formulation is that they are *decomposable* over component classifiers.

Take *redundancy* for example, given an ensemble of size k, its redundancy is $\mathcal{I}(X_{1:k}) = \sum_{i=1}^{k} I(X_i; X_{1:i-1})$. Then, if a new classifier X_{k+1} is added, the new redundancy becomes $\mathcal{I}(X_{1:k+1}) = \sum_{i=1}^{k+1} I(X_i; X_{1:i-1})$, and the only difference is the mutual information $I(X_{k+1}; X_{1:k})$. That is, during the ensemble construction process, each classifier X_i can be characterized by the following measurements:

– **Relevance**: $I(X_i; Y)$ which measures its relevance to the class label, and bounds its prediction error;
– **Redundancy**: $I(X_i; X_{1:i-1})$ which measures the dependency between current classifier and existing classifiers;
– **Conditional Redundancy**: $I(X_i; X_{1:i-1} \mid Y)$ which measures the conditional dependency between current classifier and existing classifiers given the class label;
– **Diversity**: Which is the difference between the conditional redundancy and redundancy, and measures its contribution to the ensemble diversity.

Next, we study the relationship between Eqs. 10 and 7; that is, the relation between multi-information diversity and Brown's result.

Lemma 2. *Given a set of variables $V = \{X_1, \cdots, X_n\}$, it always holds that*

$$\sum\nolimits_{k=2}^{n} \sum\nolimits_{T_k \subseteq V} I(\{T_k\}) = \mathcal{I}(X_{1:n}). \tag{11}$$

Proof. It has been shown in [1] that the *interaction information* can be expanded as a sum of entropies, i.e.,

$$I(\{X_{1:n}\}) = -\sum\nolimits_{T \subseteq V} (-1)^{|V \setminus T|} H(T), \tag{12}$$

where $V = \{X_{1:n}\}$ and $\sum_{T \subseteq V}$ denotes a sum over all possible subsets of V. Let

$$\Gamma_k = \sum\nolimits_{T_k \subseteq V} H(T_k) \quad \text{and} \quad \Theta_k = \sum\nolimits_{T_k \subseteq V} I(\{T_k\}) \ .$$

Substituting Eq. 12 into Θ_k, we can obtain $\Theta_2 = C_{n-1}^1 \cdot \Gamma_1 - \Gamma_2$ and $\Theta_3 = -C_{n-1}^2 \cdot \Gamma_1 + C_{n-2}^1 \cdot \Gamma_2 - \Gamma_3$, and in more general case,

$$\Theta_k = (-1)^k \cdot C_{n-1}^{k-1} \cdot \Gamma_1 + (-1)^{k-1} \cdot C_{n-2}^{k-2} \cdot \Gamma_2 + \cdots + (-1) \cdot \Gamma_k \ . \tag{13}$$

Substituting Eq. 13 into $\sum_{k=2}^{n} \Theta_k$ which is the left-hand side of Eq. 11, it is easy to find that the coefficient of Γ_i is

$$-\sum\nolimits_{j=1}^{n-i} C_{n-i}^{j} (-1)^j = -(1-1)^{n-i} = 0, \quad \text{for} \ \ i = 2, \cdots, n-1$$

and the coefficients of Γ_1, Γ_n are 1 and -1, respectively. Consequently, it follows that

$$\sum\nolimits_{k=2}^{n} \Theta_k = \Gamma_1 - \Gamma_n = \sum\nolimits_{i=1}^{n} H(X_i) - H(X_{1:n}) = \mathcal{I}(X_{1:n}) \ ,$$

which completes the proof. □

Corollary 1. *Eq. 7 and Eq. 10 are mathematically equivalent.*

Proof. It is obvious that the relevance terms in Eqs. 7 and 10 are the same. Based on Lemma 2, we have $\mathcal{I}(X_{1:m}) = \sum_{k=2}^{m} \sum_{T_k \subseteq S} I(\{T_k\})$. Since $I(X_{1:m}; Y)$ appears in the left-hand sides of both equations, it follows that $\mathcal{I}(X_{1:m} \mid Y) = \sum_{k=2}^{m} \sum_{T_k \subseteq S} I(\{T_k\}|Y)$. Consequently, we can reach the conclusion that two equations are equivalent. □

Hence, although our formulation is simpler, it is mathematically equivalent to Brown's formulation. Moreover, Brown's formulation decomposes the ensemble diversity over different orders of interaction, while our formulation decomposes over the component classifiers. Based on our formulation we have an approximation for estimation, which will be presented in the next section.

4 Approximate Estimation

For estimating the multi-information diversity and its related terms, one straightforward approach is to estimate the joint probability. However, the exponential number of possible variable values will make the estimation of joint distribution infeasible in practice. So we take an approximation.

For redundancy, our task is reduced to estimate $I(X_i; X_{1:i-1})$ for all i's. Here, rather than estimating the joint probabilities, we approximate $I(X_i; X_{1:i-1})$ by

$$I(X_i; X_{1:i-1}) \approx \max_{\Omega_k} I(X_i; \Omega_k) \ , \tag{14}$$

where $\Omega = \{X_{i-1}, \cdots, X_1\}$, and Ω_k is a subset of size k.

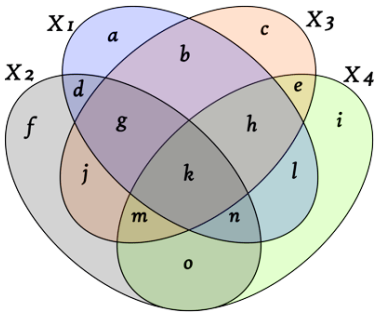
As an illustrative example, Fig. 1 depicts a Venn diagram for four variables, where the ellipses represent the entropies of different variables, while the mutual information can be represented by the combination of regions in the diagram. As shown in the right side of the figure, it can be found that the high-order component $I(X_4; X_3, X_2, X_1)$ shares a large intersection with the low-order component $I(X_4; X_2, X_1)$, where the only difference is the region e. Note that if X_1, X_2 and X_3 are strongly correlated, it is highly likely that the uncertainty of X_3 is covered by X_1 and X_2, that is, the regions c and e would be very small. Thus, $I(X_4; X_2, X_1)$ provides an approximation to $I(X_4; X_3, X_2, X_1)$. Such a scenario often happens in ensemble construction since the component classifiers generally have strong correlations.

With respect to the conditional redundancy and multi-information diversity, we take similar strategies as follows.

$$I(X_i; X_{1:i-1} \mid Y) \approx \max_{\Omega_k} I(X_i; \Omega_k \mid Y) \tag{15}$$

$$I(X_i; X_{1:i-1} \mid Y) - I(X_i; X_{1:i-1}) \approx \max_{\Omega_k} [I(X_i; \Omega_k \mid Y) - I(X_i; \Omega_k)] \tag{16}$$

To accomplish the approximation in Eqs. 14, 15 and 16, an enumeration over all the Ω_k's is desired. In this way, however, for every i we need estimate $I(X_i; \Omega_k)$ and $I(X_i; \Omega_k \mid Y)$ for C_{i-1}^k number of different Ω_k's. When k is near $(i-1)/2$, the number will be large, and the estimation of $I(X_i; \Omega_k)$ and



$I(X_4; X_3, X_2, X_1) = e + h + k + l + m + n + o,$
$I(X_4; X_2, X_1) = h + k + l + m + n + o,$
$I(X_4; X_3, X_1) = e + h + k + l + m + n,$
$I(X_4; X_3, X_2) = e + h + k + m + n + o.$

Fig. 1. Venn diagram of an illustrative example

$I(X_i; \Omega_k \mid Y)$ will become difficult. Therefore, we need to take a trade-off. In our experiments, we restrict k to be one or two.

If considering only pairwise interactions, our method (MTI) and Brown's method [2] estimate $\sum_{i=2}^{n} I(X_i; X_{1:i-1})$ by

$$\sum\nolimits_{i=2}^{n} \max_{k<i} I(X_i; X_k) \quad \text{and} \quad \sum\nolimits_{i=2}^{n} \sum\nolimits_{j=1}^{i-1} I(X_i; X_j) \ ,$$

respectively. In other words, $I(X_i; X_{1:i-1})$ is approximated by $\max_{k<i} I(X_i; X_k)$ and $\sum_{j=1}^{i-1} I(X_i; X_j)$, respectively, by our method and Brown's method.

Take $I(X_4; X_3, X_2, X_1)$ in Fig. 1 for example. It is easy to get that,

$$\begin{aligned} &I(X_i; X_{1:i-1}) - \max_{k<i} I(X_i; X_k) = I(X_4; X_{1:3}) - \max_{k<4} I(X_4; X_k) \\ &= (e + h + k + l + m + n + o) - \max\{(h + k + l + n), (k + m + n + o), \\ &(e + h + k + m)\} = \min\{(e + m + o), (e + h + l), (l + n + o)\} \ , \end{aligned} \tag{17}$$

$$\begin{aligned} &\sum\nolimits_{j=1}^{i-1} I(X_i; X_j) - I(X_i; X_{1:i-1}) = \sum\nolimits_{j=1}^{3} I(X_4; X_j) - I(X_4; X_{1:3}) \\ &= (h + k + l + n) + (k + m + n + o) + (e + h + k + m) - (e + h + k + \\ &l + m + n + o) = 2k + h + m + n \ . \end{aligned} \tag{18}$$

Note that the right-hand sides of both Eqs. 17 and 18 are nonnegative, which implies that our approximation is a lower bound never larger than the true value, while Brown's estimation is an upper bound never smaller than the true value. Moreover, it is easy to get that, at least when the following equation holds,

$$3k + h + m + n > e + l + o \tag{19}$$

the right-hand side of Eq. 18 is larger than that of Eq. 17, which means that our approximation is closer to the true value than Brown's estimation. It is worth noting in Eq. 19 that the region k represents the uncertainty shared by all the four variables, the regions h, m and n represent those shared by three variables,

Table 1. Evaluation on synthetic data

	Ground-truth	**Brown$_1$**	**Brown$_2$**	**MTI$_1$**	**MTI$_2$**
$\mathcal{I}(X_{1:5})$	1.9485	2.9418	1.9946	1.6019	1.9485
Relative error	—	50.98%	2.37%	-17.79%	0.00%

while the regions e, l and o represent those shared by two variables. This discloses that when the variables are highly correlated (such as when Eq. 19 holds), our approximation is expected to be closer to the ground truth. In ensemble construction, the component classifiers are generally highly correlated, so we expect that our approach is better for estimating ensemble diversity and related terms in practice. This will be empirically verified in the next section.

5 Experiments

5.1 Synthetic Data

To evaluate our proposed approach, experiments on synthetic data is performed at first since we can get the ground-truth information.

The task is to estimate $\mathcal{I}(X_{1:n})$. The synthetic data is generated as follows. Assume a and b are integers sampled from the interval $[1, 100]$, the variables

$$X_1 \equiv a \pmod 2, \quad X_2 \equiv a + b \pmod 2, \quad X_3 \equiv a - b \pmod 2,$$
$$X_4 \equiv a \times b \pmod 2, \quad X_5 \equiv a^2 + b^2 \pmod 2$$

are correlated. It is easy to get $p(X_i)$'s, $p(X_i, X_j)$'s, $p(X_i, X_j, X_k)$'s and $p(X_{1:5})$, based on which the ground-truth $\mathcal{I}(X_{1:5})$ can be obtained analytically.

We evaluate our method (MTI), where k is restricted to be 1 and 2, denoted by MTI$_1$ and MTI$_2$, respectively. We also evaluate Brown's method [2]. Since MTI$_1$ and MTI$_2$ consider the pairwise and three-order interactions, for a fair comparison, in addition to the method reported in [2] which considers only pairwise interactions, we extended Brown's method to consider three-order interactions by using Eq. 12 to help estimate the three-order interaction information. These two versions are denoted by Brown$_1$ and Brown$_2$, respectively. The estimated as well as the ground-truth multi-information are shown in Table 1, where the *relative error* is the difference between the ground-truth value and the estimated value divided by the ground-truth value.

From Table 1 we can find that Brown$_2$ and MTI$_2$ perform much better than Brown$_1$ and MTI$_1$, and MTI$_2$ reaches the ground-truth value. This is easy to understand since Brown$_2$ and MTI$_2$ explores the three-order interactions while Brown$_1$ and MTI$_1$ consider only the pairwise interactions. Comparing MTI$_1$ (MTI$_2$) with Brown$_1$ (Brown$_2$), it can be found that the estimation of MTI is more closer to the ground-truth, although they consider the same pairwise (or three-order) interactions. This verifies our argument that MTI is more accurate if the variables are highly correlated.

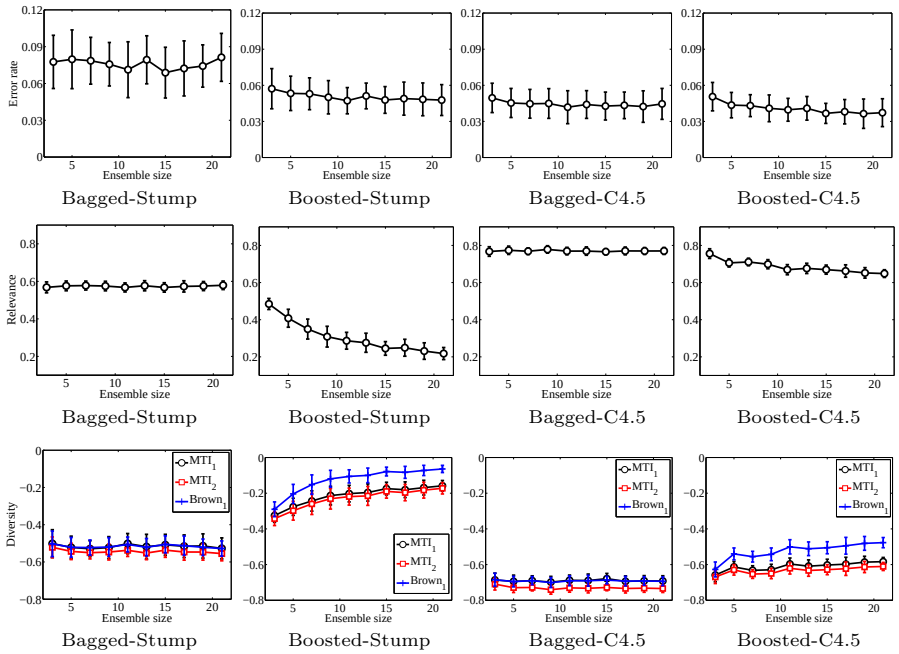


Fig. 2. *Breast Cancer* (1st row: Error rate; 2nd row: Relevance; 3rd row: Diversity)

5.2 Real Data

Next we apply our MTI approach and Brown's method to study the behavior of AdaBoost and Bagging on real data.

In the experiments we have used two types of base classifiers, i.e., *decision stump* and *C4.5 decision tree*. The ensemble sizes are set to 3 to 21. For each configuration, we execute 50 runs of hold-out tests on UCI data set *Breast Cancer* and *Heart Disease*. In each run, two thirds data are used for training while the remaining for testing. The relevance, multi-information diversity, redundancy and conditional redundancy terms are estimated on the training data, and used to explain the ensemble prediction error on test data. Here, MTI_1, MTI_2 and Brown_1 are evaluated.[2] Note that the estimated relevance, redundancy and conditional redundancy are monotonically increasing as the ensemble size increases.

[2] In the experiments on synthetic data, the task is to estimate the interaction information and we can use Eq. 12 to help estimate the three-order interaction information for Brown_2. On the real data, for estimating the diversity, Brown_2 requires to calculate three-order conditional interaction information, yet the calculation is not straightforward. Moreover, Brown_1 averages all the C_n^2 number of pairwise redundancy (and conditional redundancy) terms, while Brown_2 needs to deal with C_n^2 pairwise terms and C_n^3 three-order terms (in our approach there are always only n terms no matter what order is considered), and it is not clear whether the terms of different orders should be averaged together or not. So, in the experiments on real data we have not evaluated Brown_2.

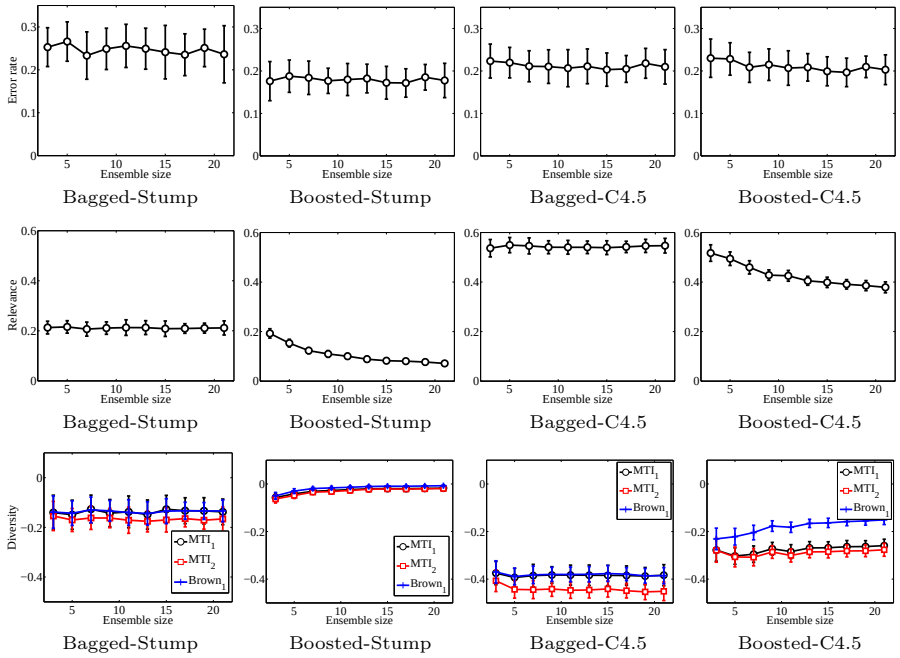


Fig. 3. *Heart Disease* (1st row: Error rate; 2nd row: Relevance; 3rd row: Diversity)

In the following, the mean of $\max_{\Omega_k} I(X_i; \Omega_k)$'s and $I(X_i; X_j)$'s are reported for MTI and Brown_1, respectively. The estimated values are averaged over 50 runs, and the mean values and standard deviations are reported. The results are shown in Figs. 2 and 3, respectively.[3]

We can see that Adaboost decreases the relevancy of its classifiers while increasing the diversity. Bagging has a very different behavior. It maintains almost constant relevancy and diversity. This is accordance with the sequential and parallel generation style of the component classifiers in AdaBoost and Bagging. With respect to ensembles with different base classifiers, it is easy to find that ensembles with decision stumps have lower relevance but higher diversity, while ensembles with C4.5 decision trees have higher relevance but lower diversity. This is easy to understand because decision stumps have lower accuracy and is more sensitive to data samples than C4.5 decision trees.

Comparing MTI_1, MTI_2 and Brown_1, it can be found that in general, all the three methods make similar diversity estimations except that Brown_1 is a little bit divergence on AdaBoost; this suggests that all these methods are useful for measuring ensemble diversity. Moreover, it can be found that the estimated values of Brown_1 is never smaller than that of MTI_1; this is accordance with our argument that MTI estimation lower bounds while Brown's estimation upper bounds the true value.

[3] The redundancy and conditional redundancy are not shown due to space limitation.

6 Conclusion

In this paper, we propose a formulation of ensemble diversity from the view of *multi-information*. This formulation is mathematically equivalent to the previous information theoretic diversity formulation [2], but is simpler and decomposable over component classifiers. Based on the formulation, we present an approximation for estimation. Experimental results show that the approximation can be used to study the behavior of ensemble methods to some extent in practice.

Acknowledgement: The authors want to thank anonymous reviewers for their helpful suggestions. This work was supported by NSFC (60635030, 60721002), JiangsuSF (BK2008018), 973 Program (2010CB327903) and the Startup Foundation in School of Mathematical Sciences at Soochow University.

References

1. A. J. Bell. The co-information lattice. In *Proc. 4th Intl. Symp. Independent Component Analysis & Blind Signal Separation*, pages 921–926, 2003.
2. G. Brown. An information theoretic perspective on multiple classifier systems. In *Proc. 8th Intl. Workshop Multiple Classifier Systems*, pages 344–353, 2009.
3. G. Brown. A new perspective on information theoretic feature selection. In *Proc. 12th Intl. Conf. Artificial Intelligence and Statistics*, pages 49–56, 2009.
4. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.
5. T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40:139–157, 2000.
6. R. Fano. *Transmission of Information: Statistical Theory of Communications*. Wiley, 1961.
7. G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19:699–707, 2001.
8. M. Hellman and J. Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Trans. Information Theory*, 16:368–372, 1970.
9. T. K. Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20:832–844, 1998.
10. L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 2003.
11. L. I. Kuncheva, C. J. Whitaker, C. Shipp, and R. Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, 6:22–31, 2003.
12. W. McGill. Multivariate information transmission. *IEEE Trans. Information Theory*, 4:93–111, 1954.
13. N. Slonim, N. Friedman, and N. Tishby. Multivariate information bottleneck. *Neural Computation*, 18:1739–1789, 2006.
14. M. Studeny and J. Vejnarova. The multi-information function as a tool for measuring stochastic dependence. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 261–298. Kluwer, 1998.
15. S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4:66–82, 1960.