Some Fundamental Issues in Ensemble Methods

Wenjia Wang, *IEEE* member

Abstract — **The ensemble paradigm for machine learning has been studied for more than two decades and many methods, techniques and algorithms have been developed, and increasingly used in various applications. Nevertheless, there are still some fundamental issues remaining to be addressed, and an important one is what factors affect the accuracy of an ensemble, and to what extent they do, which is thus taken as the main topic of this paper. The factors studied include the accuracy of individual models, the diversity among the individual models in an ensemble, decision-making strategy, and the number of the members used for constructing an ensemble. This paper firstly describes the conceptual and theoretical analyses on these factors, and then presents the possible relationships between them. The experiments have been conducted by using some benchmark data sets and some typical results are presented in the paper.**

I. INTRODUCTION

An ensemble in the context of machine learning can be broadly defined as a machine learning system that is constructed with a set of individual models working in parallel and whose outputs are combined with a decision fusion strategy to produce a single answer for a given problem. The models can be classifiers, predictors or filters, depending on the type of task – classification, prediction, regression or clustering, that the ensemble is designed to do. The rational behind the ensemble approach is based on the bare fact that no individual models can be perfectly developed for solving non-trivial real world problems.

It is common nowadays to employ some inductive machine learning algorithms to induce models, e.g. decision trees, neural nets or other models, from data quickly at a relatively low cost to build ensembles. Based on the mechanisms of construction and operation, the performance of an ensemble can be evaluated in terms of complexity, reliability and accuracy. Complexity is concerned with the computational time and memory space required and can be measured in the usual ways, but is a not major issue because computing power and resources can generally cope with most applications except extremely large and complex problems, thus it will be not covered in this paper. Reliability of an ensemble is about how reliable are the answers produced by ensembles, and may be measured by the probability that r models chosen randomly from an ensemble fail or succeed on randomly selected test data. However, in practice, it is the accuracy that people are most interested in, and achieving a higher accuracy is, in fact, the main motivation for using ensemble methods. This paper therefore will focus on investigating how the accuracy of an ensemble is influenced by what factors, and the extent of their influence.

The rest of this paper is organized as follows. The next section describes the representation of a generic ensemble, the factors that may influence the accuracy of an ensemble, and briefly summarizes the related work. Section III describes the diversity measures. Section IV: the accuracy of individual models; Section V: the number of models in ensemble, and Section VI; the experiments and results. The summary and conclusion are given in the final section.

Manuscript received on 12/12/2007. This work was partly supported by the grants from EPSRC (GR/86041/01) and ESRC (RES-000-22-0874).

W. Wang is a senior lecturer with the School of Computing Sciences, University of East Anglia, Norwich UK (wjw@cmp.uea.ac.uk).

II. ENSEMBLE METHODS

A. Generic inductive ensemble

As mentioned before, a generic ensemble can be simply viewed as a computing system built with N individual models trained from data by using one or several machine learning algorithms. Without losing generality, for a given problem with an unknown target function $F(\mathbf{x})$, a model m_i can be trained by a learning algorithm to approximate $F(\mathbf{x})$. Then, we can represent this implementation of model m_i as a function f_i that estimates $F(\mathbf{x})$:

$$F(\mathbf{x}) = f_i(\mathbf{x}) + \varepsilon_i(\mathbf{x}) \tag{1}$$

where, $\mathbf{x} = [x_1\ x_2\ \ldots x_n]$ denotes the input vector of n variables and ε_i is the error between the target and the output of m_i.

N models $\{m_1, m_2, \ldots m_N\}$ need to be generated, with some variations, to give N approximated functions $\{f_1, f_2, \ldots f_N\}$ respectively for building an ensemble V. These models can work in parallel when presented simultaneously with unseen data and each produces its own output independently. These individual outputs are combined by a decision fusion strategy to produce the final output $f(\mathbf{x},V)$ of the ensemble. *Voting* and *Averaging* are two commonly used decision fusion strategies. The averaging strategy suits the system in which the outputs of models are continuous, i.e.

$$f(\mathbf{x},V) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}, m_i)\,. \tag{2}$$

Voting strategy is appropriate for the models whose outputs $f(\mathbf{x},v)$ are categorical for classification problems and should be an index of K possible classes. The ensemble system decision can be determined by:

$$\begin{aligned} f(\mathbf{x},V) &= voting\{m_1, m_2, \ldots m_N\} = \arg\max\{c_1, c_2, \ldots c_K\} \\ c_j &= \sum_{i=1}^{N} g(f_i(\mathbf{x}, m) = j) \quad \forall\ j = 1\ to\ K \end{aligned}\,. \tag{3}$$

978-1-4244-1821-3/08/$25.00©2008 IEEE

Where c_j is the count of the output of N models for the jth class, and $g_j()=1$ if $f(\mathbf{x},m)=j$, $g()=0$ otherwise. The number of models in an ensemble, N, should be set to an odd number to avoid ties when the voting strategy is used. The function *argmax* returns a class label that has the largest value among $\{c_1, c_2, \ldots, c_K\}$.

A. Analysis on Ensemble Performance

The performance of ensemble can be evaluated from different points of views depending on the objective of the study, and this paper focuses on the accuracy or error of ensemble and the factors that may influence the accuracy or error. The concept of decomposition of the classification error in the similar manner used in liner regression error analysis was introduced to decompose the error of a model into bias and variance [1]. The *bias* term is a measure of closeness between the function $f(x)$ represented by a classifier model and the target function $F(x)$; and the *variance* term measures the output difference between the models in an ensemble, which may be viewed a kind of diversity among the models. These two terms are useful in evaluating the performance of learning algorithms used for inducing models. But, when applied in the real world, their effectiveness appears to be limited by the fact that the target function is unknown for most data-defined problems. Besides, they only measure the error of the ensemble and differences among its member models, but unable to give any information on what factors may cause these errors and variance and how. Being able to measure error is important but knowing the causes is even more important since once we know what factors, in what way (either positive or negative) and to what extent, influence the accuracy of ensemble, we will be able to work on these causes to improve the accuracy. It is therefore essential to carry out further investigations in this regard.

B. Influencing Factors

Theoretically, the accuracy of an ensemble V, denoted by $acc(V)$, can be influenced by the factors involved in the construction and operation of an ensemble, including N the number of models in the ensemble, $acc(m_i)$ - the accuracy of the individual models m_i (for $i=1$ *to* N), D - diversity among the models, S - the decision fusion strategy used in the ensemble. In general, the relationship between the accuracy of ensemble and these factors can be denoted by a function $f()$ below:

$$acc(V) = f\left(acc(m_i)\{\forall i = 1 to N\}, D, S, N\right) \quad (4)$$

As stated earlier, the aim of this work is to investigate the effects these factors have on the accuracy of an ensemble. However, this is not an easy task because these factors are tightly coupled with each other, which means, whilst trying to analyze and quantify the impact of one particular factor on ensemble accuracy, the other factors cannot be excluded from the analysis and they themselves may be affected, and their changes can affect the ensemble's performance and therefore trigger a chain of interactions. This inter-influence nature makes the analysis on all the factors simultaneously as a whole extremely difficult. Thus, some strategies and tactics must be devised in investigation. One simple strategy is that, even when these factors are inseparable in the operation of an ensemble, the influence from one factor may be reduced to minimum or kept as constant (known or unknown) while investigating another.

In this study, three specific tactics are devised to explore (1) the relationships between the ensemble accuracy and diversity among the individual classifiers, $acc(V)=f(D \mid (acc(m_i), N, S))$, given that accuracy of individual models $acc(m_i)$ is known or bound at the lower end, and N and S are fixed; (2) the relationships between the ensemble accuracy and the accuracy of individual models, assuming that N, D and S are known and constant or limited to a certain range, i.e. $acc(V)=f(acc(m_i)|(N,D,S))$; (3) the relationship between the ensemble accuracy and the number of models used in an ensemble, assuming that the other factors: accuracy of individual classifiers $acc(m_i)$, diversity D and decision fusion strategy S, are known or fixed ideally, i.e. $acc(V) = f(N \mid (acc(m_i),D, S))$. The experiments were carried out by using data sets from the UCI data repository but only some results are presented in the paper due to the limit of space.

C. Related Work

Many techniques and algorithms have been developed to build ensembles for different applications. Bagging [2], Boosting [3], Adaboost [4], Wang's Hybrid ensemble [5, 6] of neural nets and decision trees, are just some of relatively popular methods devised with a consideration of introducing diversity into ensemble. Breiman's Random Forests [7] attempts to introduce diversity into ensemble by selecting the features at random when inducing decision trees.

The performance of various ensembles has also been studied by many researchers [6, 8, 9, 10, 11 and 12]. The findings from the previous studies can be simply summarized as that an ensemble will be beneficial when its member models have accuracy more than a random guess and are diverse enough from each other. However, several critical issues remain open including: (i) of many existing definitions of diversity measure, which one is more effective? And for which definition and at what extent a diversity value is enough? (ii) How reliable is the answer produced by an ensemble and how to estimate the reliability of an ensemble? (iii) What are the possible relationships between ensemble accuracy and the other factors above mentioned. This paper attempts to draw some more research attentions to these issues by exploring some fundamental aspects related to them.

III. ENSEMBLE DIVESITY

Diversity is usually considered as a quantified estimate of the difference of making the same errors among models in an ensemble. There are many diversity definitions and ways to evaluate diversity. One type of diversity is estimated through probability analysis on the N models (in an ensemble) that disagree simultaneously in decision-making.

There are as many as dozens of different definitions of diversity proposed from different points of view and formulations. Basically they can be classified into two categories in terms of the difference measured between the

models, pair-wised and non-pair-wised. The effectiveness of some common ones were summarized in [13] and further evaluated by Bian and Wang [14]. One of their conclusions is common, that is, because the pair-wised diversity measures only consider the difference between two models, the pair-wised definitions are not effective in measuring diversity and show no or little relation with the accuracy of the ensemble and thus are not useful.

In practice people are more interested to know, given an ensemble of N models, what is the probability that all the models fail on test data coincidentally. Partridge et al [10] defined this probably as the Coincident Failure Diversity (CFD) and pointed out that there can be two types of coincident failure diversity, i.e. the CFD among the models in an ensemble, named the intra-Coincident Failure Diversity, and the diversity between the ensembles, named the inter-ensemble diversity. As they will be used and evaluated in this study we give their definitions. It will also be useful to give a brief description on the reliability measures of ensemble as they will be used in formulating the diversity measures.

A. *Ensemble reliability*

Ensemble reliability can be considered as the probability that any randomly chosen models from a given ensemble of N models (or classifiers) give a correct answer on randomly selected test data. We can estimate this probability by calculating the failure probability of a given number of models drawn from an ensemble. Specifically, we can work it out in the following manner.

Firstly, we define q_n the number of examples (from a test data set of Q samples) that fail on exactly n models in ensemble V. Then the probability p_n - exactly n models fail on randomly selected test data can be calculated by:

$$p_n = \frac{q_n}{Q}, \qquad n = 1,...., N \tag{5}$$

Then, the probability that r randomly chosen classifiers fail on a randomly chosen input can be estimated by:

$$p(r) = \sum_{n=1}^{N} \frac{n}{N} \frac{(n-1)}{(N-1)} \cdots \frac{(n-r+1)}{(N-r+1)} p_n \tag{6}$$

So the probability that r models succeed on a randomly chosen data sample will be $1 - p(r)$.

Furthermore, the stability of an ensemble can be represented by the probability that any q models taking from a subgroup of n models from a given ensemble will produce correct answer if the simple voting strategy is used in determining the decision. For example, given a subgroup of 3 models from randomly an ensemble of N models, we may want to know the probability that 2 out of these three will give the right answer, or the probability that 1 out of 3 gives the right answer. They can be simply denoted $p(2/3)$ and $p(1/3)$ respectively.

B. *Intra-ensemble Coincident–Failure Diversity (CFD)*

For a given ensemble of N-models, the CFD is defined as the probability that the models fail simultaneously.

$$CFD = \begin{cases} \dfrac{1}{1-p_0} \sum\limits_{r=1}^{N} \dfrac{N-r}{N-1} p_r, & if\ p_0 < 1 \\ 0, & if\ p_0 = 1 \end{cases} \tag{7}$$

CFD$\in$[0, 1]. When CFD = 0, it indicates either that the failures are the same in all the models - hence there is no diversity, or no test failure at all, i.e. all the models are perfect and identical - hence no diversity (in this case there is no need for diversity because a perfect model has been found). CFD = 1 when all the test failures are unique to one model, the ensemble is perfect and always produce the correct answer when the majority-voting strategy is applied.

C. *Inter-Ensemble Diversity*

In ensemble applications, sometimes many ensembles may be constructed with different variations. Then it is helpful to know how much difference exists between the ensembles, or the models from different ensembles that fail on a randomly selected input simultaneously. Extending the earlier CFD measure to two ensembles A and B, the coincident failure diversity CFD_{AB} [10] between A and B can be defined as

$$\begin{aligned} CFD_{AB} = & \frac{1}{1-p_{00}} \left(\sum_{i=1}^{N_A} p_{i0} + \sum_{j=1}^{N_B} p_{0j} \right) \\ & + \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \left(\frac{i(N_B - j)}{N_A N_B} + \frac{j(N_A - i)}{N_A N_B} \right) p_{ij} \end{aligned} \tag{8}$$

Where p_{ij} is the joint probability that i models from A and j models from B simultaneously fail on a randomly chosen input. $CFD_{AB} \equiv 0$ when p_{00}=1 or p_{NANB}=1. CFD_{AB} = 1 when the two ensembles are said to have maximum diversity. The measure can be extended to multiple ensembles.

D. *Minority-failure diversity (MFD)*

However, based on our research, we believe that an effective diversity measure must be defined in relation to the decision-making strategy of the ensemble system. For example, if one ensemble uses *averaging* as its decision fusion strategy (i.e. the ensemble's output is determined by calculating the average of all the individual outputs of member models). In such a case, there is no point in measuring the diversity between the member models as the final output is always the average of the member models' outputs. Conversely, if one uses the simple-majority voting method as the ensemble's decision fusion strategy, in such a case, it will be very vital to know the probability that the majority of the models in an ensemble do not make the same error coincidently. So, we believe that a more effective diversity measure should somehow reflect the difference in this regard. We therefore introduced a diversity measure that estimates a kind of difference exists among the models of an ensemble in the

right places so that only a minority of the models fails on the given test data and the majority produces the right answer. Thus when the simple-majority voting strategy is applied, the ensemble is able to take the answer of majority models as its final decision. So, the majority-correct probability or minority-failure diversity (MFD), can be defined as:

$$MFD = \sum p(\text{minority models fail}) = \sum_{r=0}^{(N-1)/2} p_r \qquad (9)$$

Compared with the non pair-wised diversity definitions, this measure is intuitively easier to understand and interpret as it is derived directly from a commonly used decision making strategy, i.e. simple majority voting and is also mathematically simple to compute.

It is generally acknowledged that diversity among the models plays a key role in making the ensemble approach beneficial when the individual models are less than perfect. If there is no diversity between the member models in an ensemble, it will not improve accuracy and then there will be no need to build such ensembles. On the other hand, if the ideal diversity is somehow achieved among the member models in an ensemble, this ensemble shall always produce the correct answer. However, in reality, it will be extremely difficult to achieve the ideal or maximum diversity. So, given the facts that (1) no perfect individual models can be generated and (2) no maximum diversity achieved, then the critical questions that should be asked include: what accuracy of individual models and what diversity value are sufficient to make a better ensemble? These issues will be analyzed and discussed in the next sections.

IV. ACCURACY OF INDIVIDUAL MODELS

When considering the influence of the accuracy of individual models on the accuracy of an ensemble, we can conceptually estimate it with respect to several different conditions classified simply as: all the models have identical or similar accuracy and the lower bound of their accuracy is higher than the default accuracy, or the higher bound of their accuracy is lower than the default accuracy.

A. *The lower bound of accuracy*

As researchers have pointed out, when the accuracy of the individual models in an ensemble is better than a random guess, the ensemble should be able to improve the accuracy if these models are diverse enough from each other. This random guess or default accuracy can then be used as the accuracy lower bound to the individual models when selecting models for building an ensemble. The accuracy lower bound *acclb* for a classification task varies with the number of target classes of problem and can then be estimated by:

$$acclb(m) = \lim_{N\to\infty} \frac{N+1}{KN} = \frac{1}{K} \qquad (10)$$

Where,. The relationship between K - the number of the target classes, and the accuracy lower bound is shown in Figure 1. For dichotomous classification problems, where K=2, *acclb* = 0.5.

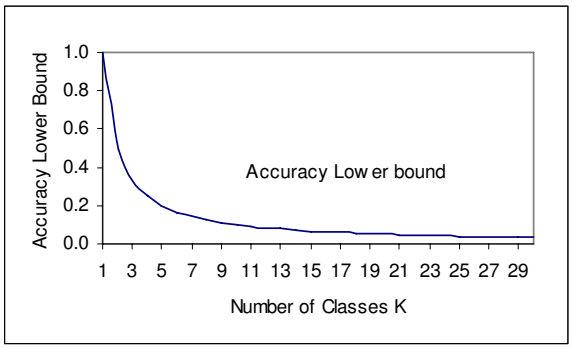

Figure 1. The accuracy lower bound for individual models with respect to the number of target classes.

B. *When acc(m_i) >= accuracy lower bound acclb*

When the lowest accuracy of the individual models in an ensemble is higher than or equal to the lower bound *acclb* defined as above, and all the others have the same or slightly higher accuracy, then the accuracy of the ensemble will be determined by three factors: diversity D, the number of the models N and decision fusions strategy S.

In such a case, when the simple majority voting strategy is applied, the range of the accuracy of the ensemble shall be bounded by *acclb <= acc(V) <= 100%,* and the actual accuracy of the ensemble will also be influenced by the diversity and number of the models in the ensemble.

When the models are identical, it is the worst case because there is no diversity among them at all, i.e. D=0. Thus, if one model makes a mistake all the others make the same mistake as well. So the ensemble acts the same way as any individual model does, the number of models becomes irrelevant, and the ensemble's accuracy is equal to any individual one's, *acc(V) = acc(mi)*.

When models are different and have some diversity, D>0, then the accuracy of the ensemble can be improved. The degree of improvement will be determined by the degree of diversity and the number of the models in the ensemble, but usually not linearly proportional.

The best case is obviously that, when the maximum diversity exists among its member models, the ensemble should be 100% accurate, given that the number of the models is sufficient. Figure 2 depicts the range of the accuracy of an ensemble that consists of 3 models (i.e. N=3) against the accuracy of individual models when maximum diversity exists. It shows that when the maximum diversity does exist, if the individual models have accuracy equal to or higher than, the majority threshold (2 out of 3, 66.7% in this case), then the ensemble will be 100% accurate. Conversely, if the accuracy of individuals is lower than the majority threshold, then the ensemble cannot possibly achieve 100% accuracy, and its accuracy will be bounded by the region shown in the figure.

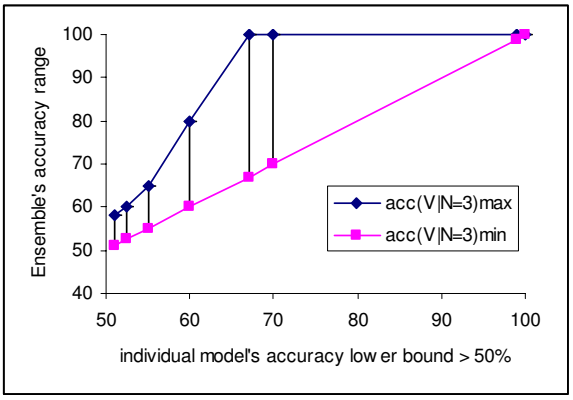

Figure 2. Accuracy range of ensemble of 3 models against the accuracy lower bound of individual models.

If one wants to improve ensemble accuracy but the accuracy of individuals has been trained to a stable region (but not over-trained), then increasing the number of models may offer some benefits. The relationship between the accuracy of the individual models and the number of models in an ensemble will be analyzed and depicted in the next section

C. *When the accuracy of individual models < acclb*

When the highest accuracy of the individual models in an ensemble is lower than the accuracy lower bound *acclb*, then the accuracy of the ensemble should be bounded by 0 <= *acc(V)* <= *2acc(m)*. The reasons are explained as follows.

In such cases, the diversity issue becomes more complicated as all the existing definitions are unable to represent and differentiate the different types of differences among the individual models. Basically, there can be two kinds of diversity existing among the models in an ensemble, which can be defined as destructive and constructive diversity, or negative and positive diversity respectively. The destructive diversity is reflected by the difference among the individual models that, when a voting strategy is employed, the number of wrong answers can become dominant and the ensemble then produces more wrong answers than any individual model. The worst scenario is that the wrong answers produced by the models may always win in voting so that the ensemble as a whole cannot produce any correct answer at all, i.e. *acc(V)*=0%, the lowest accuracy. When there is no diversity among the models, i.e. all the models are identical, the ensemble will have the same accuracy as any of individuals, i.e. *acc(V)=acc(m).*

On the other hand, when the constructive diversity exists among the models and this type of diversity can make the ensemble more accurate than individuals. The upper bound of the accuracy of ensemble will be achieved when the maximum constructive diversity exists among the models. The actual value of the upper accuracy also depends on the accuracy of individual models and will be analyzed for dichotomous classification problems below.

D. *When the accuracy of individual models < 50%*

For Boolean classification problems, the default accuracy is ½, i.e. 50%, which is set as the upper bound of accuracy of individual models in this case. Then, if the models in an ensemble *V* have maximum constructive diversity, then the highest accuracy an ensemble can achieve is bounded by:

$$\forall \ acc(m) \le 50\% \text{ and } \exists \text{ maximum and constructive D}: \quad acc(V)_{\max} = \lim_{N\to\infty} \frac{2N\,acc(m)}{N+1} = 2acc(m) \tag{11}$$

This simply indicates that the maximum accuracy of an ensemble V may theoretically reach to as much as twice the accuracy of the models in *V* iff they have ideal diversity. Figure 3 depicts the relationships between the accuracy of an ensemble and the minimum number of the models required to build an ensemble, given different accuracy lower bounds (from 50% down to 25%) of the models and ideal constructive diversity among them.

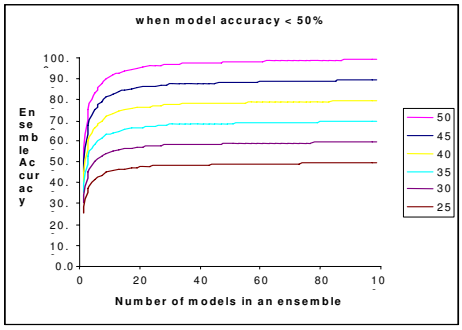

Figure 3. The relations between the accuracy of ensemble and individual models and the number of models in ensembles.

This shows that, in principle, even if the accuracy of individual models is less than the accuracy lower bound, the ensemble is still able to improve its accuracy, provided that the models are constructively diverse enough. Nevertheless, this may not have much use in practice because it is very rare that individual models cannot be trained to achieve accuracy higher than the random guess or lower bound, given that the available data are reasonably sufficient. The point is that, in case where the data are difficult to collect and the available data are not enough to train the individual models to be more accurate than the lower bound, then the ensemble approach may still be capable of improving accuracy if the models are sufficiently constructively diverse.

We have up to now presented some theoretical analysis on the influence of the accuracy of individual models on the accuracy of an ensemble and given the lower bound and higher bound of accuracy of ensemble under different conditions. Some computational experiments have been conducted in these respects and the results will be presented in the latter sections. Nevertheless, more mathematical

analysis should be conducted in the future to consolidate these findings.

V. NUMBER OF MODELS IN ENSEMBLE

This section attempts to quantify the influence of the number of the models used in building an ensemble and the accuracy of the ensemble. Many empirical studies [9, 11, 12, 15, 16] have been done and the number of models or classifiers they used in an ensemble varies from ten members to hundreds, and in some cases up to 10,000 [17] or more for bagging or boosting based ensembles. Their research did not really result in any criteria or guidelines for determining optimal number of models. It thus appears that more in-depth analysis is needed to explore how and under what conditions the number of models or ensemble size will affect the accuracy of ensembles. Again, as the operation of ensemble is complex and involves many factors mentioned earlier, attempting to analyze and quantify this effect alone is very difficult. So, the same strategy employed for the earlier analyses is used again. We assume that all the other factors such as accuracy of individual classifiers $acc(m_i)$, diversity D and decision fusion strategy S, are known and fixed ideally, or can be estimated in advance, then we can focus on investigating the relationship between the accuracy of an ensemble and the number of models, denoted by $acc(V)=f(N \mid acc(mi), D, S)$.

A. Ensemble win threshold in simple majority voting

Firstly, as in common practice, we choose the simple majority voting method as the ensemble decision fusion strategy S, it is then easier to work out the majority number for an ensemble with a given number of models. For quantifying the influence of the number of models, we introduce a term named the win threshold: $r=(N+1)/(2N)$, which is defined as the majority value of a given odd number N. Figure 4 shows the relationship between the win threshold r and N.

Obviously, for N=3, r=2/3=0.667 and $r=lim(N+1)/(2N) \Rightarrow 0.5$ when $N \to \infty$. This is well known but its relations with the accuracy of individual models and the ensemble are not clear and also depend on other factors. This win threshold should be taken into account in conjunction with the default accuracy when determining the lower bound of the accuracy of individual models. Since in most cases of application, the accuracy of individual models and their diversity can be easily estimated on training and validation data, it will be useful to predict roughly how many models are needed to construct an effective ensemble, which can be represented by $N_{min}= f(acc\{V\} \mid \text{acc(m)}, D)$.

This relationship can be established when the values of the other associated factors such as individual models' accuracy $acc(m)$ and diversity D are known and can be estimated

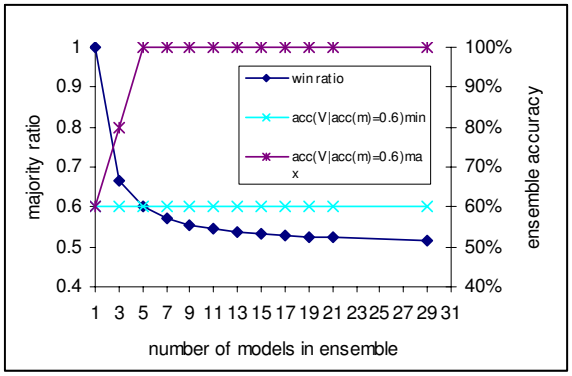

Figure 4. The win threshold of an ensemble against the number of models in the ensemble when the simple majority voting method is used. The range of accuracy of ensemble when the accuracy of the individual models is bounded at 60%.

B. Relationship $acc(V) = f(N \mid (acc(m)>=1/K, D))$

For a given N, when the accuracy $acc(m_i)>=r\%$, if the maximum diversity exists among the models, i.e. D=1, then the ensemble accuracy $acc(V)$ should be able to reach 100%. Figure 5 shows the range of the number of models required for building ensembles depending on the value of diversity among the models. For example, if set N=3, then r=0.667, then if the individual models have accuracy $acc(m)$>=66.7% and maximum diversity, the ensemble should always be correct. But if the maximum diversity cannot be achieved, which is almost certain in reality, then more models should be used and the actual number should be determined by the accuracy of individual models and the estimated value of diversity among them.

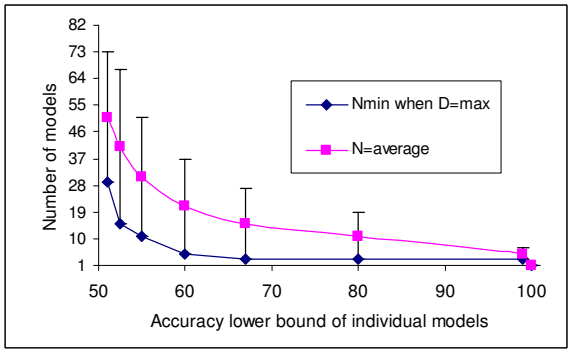

Figure 5. Number of models needed in ensembles when the accuracy of models is estimated and the empirical ranges of N when diversity is not maximum.

C. Relationship acc(V)= f(N | acc(m)<1/K, D=max)

This case has been discussed in Section IV.D and depicted by Figure 3. When $acc(m) < 1/K$, if maximum constructive diversity exists, the ensemble can perform better than individuals, if N is large enough. However, in reality, the prior probability that models have constructive diversity should be the same as having destructive diversity, it would be difficult to build beneficial ensembles when $acc(m)<1/K$, unless some appropriate strategies are devised and applied.

VI. EXPERIMENTS AND RESULTS

A. Design of experiments and data sets

The three sets of experiments described in Section II.B, (i.e. exploring the influence on the accuracy of the ensemble from diversity D, accuracy of individual models $acc(m)$, and the number of models in ensemble N) were conducted by using twelve data sets from UCI data repository. Table 1 lists the characteristics of these data sets: name, numbers of attributes, of classes, of samples and their default accuracy.

In this study, the models are decision tree classifiers, induced by using algorithm C4.5 and the ensembles are constructed by a method we modified from Random Forest, which can be briefly described as follows. Firstly, a proportion of the features are randomly selected. Then a tree is constructed using only those features. This process is repeated until the required number of decision trees has been generated. The trees are taken as member candidates for forming ensembles. The simple majority voting strategy is employed to produce the final decision of the ensemble, but the decision making strategy *averaging* is also used for comparison. The CFD is used to measure diversity among the classifiers and the minority-failure diversity for making voting decisions.

As in usual practice, each data set is partitioned into training and testing subsets with equal proportions. On the training data a 10-fold cross-validation is performed so 10 decision trees are generated. Then the data set is shuffled and repartitioned in the same manner for another run. This process is repeated for 10 times. The results given are the average of the 10 runs. Because the limit of the space, only small parts of the results are presented in the paper and more details can be made available on request.

B. Constructions of ensembles

To investigate the relationships as described earlier, one hundred decision trees were generated using the method described. From these one hundred trees thirty ensembles were created in the following ways:

- The first ten ensembles were created by selecting, one by one, the ten trees with the greatest accuracy on the training data. Thus, the first ensemble comprised only a single tree and the tenth comprised ten trees.
- The next ten ensembles were created by adding, one by one, ten randomly selected trees to the previous ten ensembles.

Table 1. The data sets used in this study.

Data	No. of attributes			no.	no.	default
	all	nom	cnt	classes	records	acc(%)
BC1	10	0	10	2	699	65.52
BC2	30	0	30	2	569	62.74
Cleve	13	7	6	2	303	54.46
CMC	9	5	4	3	1,473	42.70
CRX	15	9	6	2	690	55.51
Ecoli	8	1	7	8	336	42.56
Glass	10	1	9	7	214	32.71
Horse	22	15	7	2	368	63.04
Mush	21	21	0	2	8124	51.80
Pima	8	0	8	2	768	65.10
Sick	25	18	7	2	3163	90.74
Soy	35	35	0	19	683	13.47

nom=nominal type of attributes, Cnt = continuous attributes.

- The final ten ensembles were created by adding a further ten trees, one by one, with the worst accuracy on the training data, to the previous ten ensembles.

The idea was to generate ensembles with varying degrees of diversity so that the accuracy of these ensembles could be investigated. Using the procedure defined above for creating ensembles, the mean accuracy of the trees in the ensembles must strictly decrease as the first ten and last ten trees are added. Therefore, it is possible to identify the ensembles with 1-10 decision trees and 21–30 trees and, by default those with 11–20 trees.

C. Experimental results and discussions

Figure 6 shows the training and testing experimental results obtained for BC1 data set.

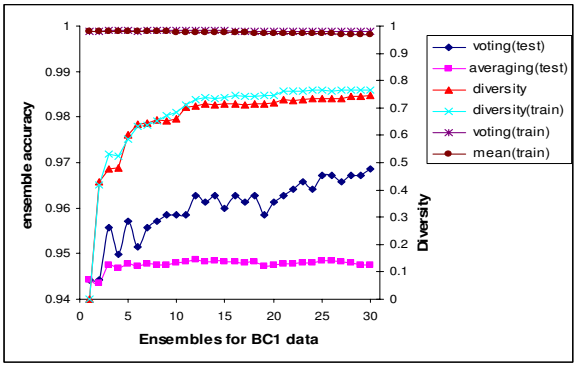

Figure 6. The influence of the diversity on accuracy of 30 ensembles by using averaging and voting methods.

It should be noted that the accuracies (of training and testing) of each ensemble, produced by averaging and voting strategies, are shown in the Figure, along with their diversity values using the diversity y-axis on the right. In this manner, through a bit complex, the influence of the diversity on the voting accuracy can be compared visually and is discussed below with respect to the earlier theoretical analyses.

(1) Regarding diversity, obviously no diversity exists in one model ensemble and very small in two-model ensembles. but when $N \geq 3$, in general, it can be seen that the diversity

increases rapidly as the number of models increases until *N* reaches to 11 or 13, after which, the diversity still increases but at a relatively small rate. This phenomenon is observed in both training and testing as shown in the plots (the middle two lines on the figure).

(2) Regarding accuracy, one characteristic that is clearly indicated in the figure is that, in training the *voting* and *averaging* strategies produced almost the identical results (the top two lines on the figure, note that these two lines are almost inseparable, thus like a thick single line) regardless of what the values of the diversity were, but on the test data, they produced very different accuracies. The bottom line is the ensemble's accuracy produced by the averaging strategy, which remains almost level and even drops a little bit at the end. However, the accuracy of ensembles, as shown by the second line from bottom, when the voting strategy is applied, keeps improving in general as the diversity increases, though it may wobble a bit (for the reason that will be discussed in the next paragraph). Even in the last ten ensembles for which the worst decision trees were added (that is why the average accuracy is reduced), the diversity is extended even further therefore the voting accuracy of ensembles is improved even further. The difference between the voting accuracy and the mean accuracy becomes increasingly larger as the diversity increases.

(3) Regarding the number of models *N*, it is clear that when the ensembles have fewer members (N<=5 or 7), which may be more accurate on the training data but less diverse, then the ensembles performed relatively bad on the test data. When *N* increases, even the less accurate decision tree classifiers are added into the ensembles, but the diversity still increases, thus the voting accuracy is improved, even though the averaging accuracy drops. Another interesting point shown from the results is the up-down wobbling of the voting accuracy as pointed out earlier. A close examination on the results found that the most of drops were actually caused when the ensembles have even numbers of members, e.g. 4, 6 and so on, and ties could happen in voting. We treated a tie as a lost in voting and the win-threshold in such cases becomes $(N+2)/2N$, instead of $(N+1)/2N$ for odd *N*. So, the win threshold becomes higher and thus voting for a win needs more votes to become a majority, e.g. when an even *N*=4, the win-threshold is 3/4=0.75, but for an odd N, e.g. *N*=3, the win-threshold is 2/3=0.667, about 0.08 lower than that of *N*=4. When N becomes bigger, the difference of the win-thresholds between the ensembles of even and odd numbers of models becomes smaller. For example, it is down to 0.016 when *N*=29 and 30, and may be neglected when *N* is even larger. This explains why that the voting accuracies of the ensembles with even numbers of members dropped quite considerably for smaller even *N*s and the wobbling amplitude becomes smaller and smaller as *N* becomes larger and larger. This finding also justifies why odd numbers of *N* should be selected when building ensembles as odd-numbered ensembles have lower wining thresholds and therefore are likely to be more accurate when the simple majority voting strategy is applied and this is particularly important when *N* is smaller.

VII. SUMMARY

This paper has analyzed the influence of the important factors involved in an ensemble's construction and operation. Specifically, it discussed three factors: diversity, individual model's accuracy and number of models in an ensemble and presented some relationships between these factors and the accuracy of ensembles under certain conditions. The experiments with some commonly used data sets have been carried out to verify the findings and the results achieved are largely in line with the theoretical analyses. We found that building ensembles with the most accurate models may not result in better ensembles and, instead, adding some less accurate models can make the ensemble more diverse and thus more reliable and accurate on test data. Another point is when voting strategy is used the ensemble with odd *N* has a lower win threshold than even *N* and is better. Further work should include in-depth mathematical analyses and empirical investigations for more common conditions in applications, such as when diversity is relatively low and model's accuracy is relatively high.

REFERENCES

[1] Geman, S. et al. (1992): Neural networks and the bias/variance dilemma. Neural computation, vol.4, pp1-58.

[2] Breiman, L.(1996): Bagging prediction. Machine learning, 24, pp123-140.

[3] Schapire et al (1997): Boosting the margin: A new explanation for the effectiveness of voting methods. In Fisher, D.(Ed) Machine Learning: Proceedings of 14th Int. Conference, pp322-330, Morgan Kaufmann.

[4] Freund, Y. & Schapire R.E. (1996): Experiments with a new boosting algorithm, in L. Saitta, ed., Machine Learning: Proceedings of the 13th national conference, Morgan Kaufmann. pp148-156.

[5] Wang, W. & Partridge, D. (1998): Multi-version neural network systems. Proceedings of neural networks and their applications, NEURAP'98, Marseilles, France, 1998, pp351-357.

[6] Wang, W. et al.. (2001): Hybrid Ensembles and Coincident-Failure Diversity, in Proceedings of IJCNN01, pp2376-2381.

[7] Breiman L. (2001): Random Forests. Machine Learning 45(1): pp 5-32.

[8] Hansen, L. et al. (1990): Neural network ensembles. IEEE Trans. Patterns Analysis and Machine Intelligence, vol. 12 pp993-1001.

[9] Optiz, D. & Maclin, R. (1999): Popular Ensemble Methods: An Empirical Study. J of Artificial Intelligence Research, 11. pp169-198.

[10] Partridge, D. et al. (1996): Engineering multi-version neural-net systems. *Neural Computation*, vol. 8, pp869-893.

[11] Wang W. et al (2000): Diversity between Neural Networks and Decision Trees for Building Multiple Classifier Systems. *Multiple Classifier Systems*: pp 240-249.

[12] Richards, G. & Wang, W. (2006): Empirical Investigations on Characteristics of Ensemble and Diversity. IJCNN06, July 2006.

[13] Kuncheva, L..(2003): Measures of diversity in classifier ensembles and their relationships with the ensemble accuracy. Kluwer Academic Publisher, Netherlands, 2003.

[14] Bian, S. & Wang, W. (2007): On diversity and accuracy of homogeneous and heterogeneous ensembles. Int. Journal of Hybrid Intelligent Systems, Vol. 4, No.2, pp103-128.

[15] Wang, W. et al.. (2001): Hybrid Ensembles and Coincident-Failure Diversity, in Proceedings of IJCNN01, pp2376-2381.

[16] Wang, W. et al. (2005). Hybrid Data Mining Ensemble for Predicting Osteoporosis Risk. IEEE-EMBS 2005. 27th Int. Conf. on Engineering in Medicine and Biology, pp886 – 889.

[17] Grove, A. and Schuurmans, D. (1998) Boosting in the limit: Maximizing the margin of learned ensembles. In Proc. of the 15th Nat. Conf. on Artificial Intelligence (AAAI-98).