Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Fusion 6 (2005) 37–48

INFORMATION FUSION

www.elsevier.com/locate/inffus

Assessing the predictive accuracy of diversity measures with domain-dependent, asymmetric misclassification costs

Mordechai Gal-Or [a,*], Jerrold H. May [b], William E. Spangler [a]

[a] A.J. Palumbo School of Business Administration, Duquesne University, Rockwell Hall, Pittsburgh, PA 15282, USA
[b] J.M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA, USA

Received 30 October 2003; received in revised form 8 April 2004; accepted 16 April 2004
Available online 21 July 2004

Abstract

We explore the relationship between diversity measures and ensemble performance, for binary classification with simple majority voting, within a problem domain characterized by asymmetric misclassification costs. Extending the work of Kuncheva and Whitaker [Machine Learning 51(2) (2003) 181], we compare a set of diversity measures within two different data representations. The first is a *direct* representation, which explicitly allows for consideration of asymmetric costs by indicating the specific values of the predictions—which in turn allows for a distinction between more costly misclassifications in this domain (i.e., actual 0 predicted as 1) and less costly ones (i.e., actual 1 predicted as 0). The second is an *oracle* representation, which indicates predictions as either correct or incorrect, and therefore does not allow for asymmetric costs. Within these representations we identified and manipulated certain situational factors, including the percentage of target group members in the population and the designed accuracy and sensitivity of each constituent model. Based on a neural network comparison of diversity measures and ensemble performance, we found that (1) diversity measure association with ensemble performance is contingent on the data representation, with Yule's Q-statistic and the coincident failure measure (CFD) as the best indicators in the direct representation and CFD alone as best indicator in the oracle representation, and (2) diversity measure association with ensemble performance varies as situational factors are manipulated; that is, diversity measures are differentially effective at different factor levels. Thus, the choice of a diversity measure in assessing ensemble classification performance requires an examination of both the nature of the task domain and the specific factors that comprise the domain.

Keywords: Multiple classifiers; Ensemble/committee of learners; Dependency and diversity; Unequal misclassification costs; Data mining

1. Introduction

In this paper, we empirically investigate the relationship between a collection of diversity measures and two performance criteria that are of interest in a two-group marketing classification application with unequal misclassification costs. The complexity of the application leads us to believe that ensemble classifiers should be more effective than individual approaches, so that if reliable predictive models linking the diversity measures to the performance criteria of ensembles can be formed, then the process of constructing effective ensembles might be simplified. Because diversity has been advocated as a positive ensemble characteristic [7,18, 23,25, 28,29], there is reason to believe that some measure(s) should be linkable to performance. Studies have explored the use of diversity measures both in assessing the performance of ensembles (e.g. [30]), as well as in determining the composition of an ensemble through 'thinning' (e.g. [2]).

Because no single method for building and using an ensemble(s) is appropriate across all situations [9], it is important to understand how situational characteristics impact the structure and performance of ensembles. Our work therefore focuses on the issues of diversity and cost

[*] Corresponding author. Tel.: +1-412-396-4002; fax: +1-412-396-4764.
E-mail address: galor@duq.edu (M. Gal-Or).

1566-2535/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2004.04.006

in the demographic classification of television viewers from their viewing behavior. The television advertising industry is confronted by the growth of digital personal video recorders (PVRs), which allow viewers to digitally record television programming, and to skim over or eliminate ‘in stream’ commercials [5]. More importantly for this research, because a PVR can be programmed, it also can record the viewing patterns of households and disclose their choices to others. Advertisers and service providers then can use data mining methods to infer characteristics of the household from its viewing choices—i.e., to build a profile of the household—and then send advertising to the PVR that is relevant to that type of household. The benefit to the viewer is that he or she receives advertising that potentially is more relevant and useful, and by implication receives fewer unwanted commercial messages. From the perspective of the advertiser, such targeted advertisements are more likely to be viewed and to influence purchasing behavior than non-targeted ones.

To deliver targeted advertising, it is necessary to be able to determine the members of a target group in a way that is accurate, unobtrusive and auditable. We developed a classification system that identifies the demographic and psychographic (behavioral) characteristics of viewers from their viewing patterns [34]. Our work is based on the premise that “you are what you watch”—i.e., an individual’s viewing habits tends to reveal his or her basic characteristics. Viewing patterns include the types of shows viewed, how often each show is watched, the time each show is watched, the duration of each viewing, etc. Data mining techniques can use those constituents to identify subsets of viewers rich in the target group. Table 1 illustrates the gains achieved from viewer profiling for five target gender/age segments defined by Nielsen Media Services, Inc. (NMSI) (see [34] for a discussion of segmentation strategies). From Table 1, if an ad is sent to everyone, 25.18% of the recipient households include a female aged 18–34. If an ad is sent only to households selected by the classification system, 58.06% of them include such a person. A household selected by the classification system is 58.06/25.18 = 2.3 times more likely to be of the desired type than one selected at random; that ratio is the model’s lift. The lifts in Table 1 are different for different segments because some groups are easier to predict based on their viewing patterns.

Thus, we found that television viewing data can be used to profile viewers. But more importantly, we also found that different data segments (e.g., weekend vs. weekday viewing) and different data mining algorithms (e.g., neural networks, logistic regression and linear discriminant analysis) generate different sets of predictions. Motivated by this discovery, and the general sense in data mining research that combining classifiers can provide better predictions than individual models [9], we have taken an ensemble approach to classification of television viewers.

A classifier or ensemble might be considered to perform ‘well’ if it simply produces better predictions than random guessing [9]. Similarly, an ensemble might be considered accurate if it produces predictions that, overall, are better than the predictions of its constituents. Evidence suggests that ensembles indeed can be more accurate than individual models [10,26]—but only when their predictions reflect some level of diversity—i.e., when they tend to disagree [1,4,8,10,15,18]. Bagging and boosting methods, for example, improve performance because they produce diverse classifiers [3,12,31]. While the conceptual meaning of diversity is agreed upon—e.g., Dietterich defines diversity as the tendency of constituent models to produce different errors on new data points [9]—multiple ways have been proposed to measure it.

From a set of 10 established diversity measures, Kuncheva and Whitaker empirically explored the degree to which the individual measures are indicative of ensemble performance [21]. In our research we extend the work of Kuncheva and her colleagues [19–22,32,35] by including more diversity measures, by building models using multiple predictor diversity measures, and by incorporating unequal misclassification costs—the latter resulting in different and multiple performance measures and hence more complex predictive models. Kuncheva and Whitaker considered several pairwise and groupwise measures of diversity between and

Table 1
Performance of a viewer classification system, where h means that a household includes at least one member of the demographic group and $\hat{h}$ means that the data mining system predicts that the household contains at least one such person

Demographic group	Prob(h)	Prob($h\|\hat{h}$)	Lift $= \frac{\text{Prob}(h\|\hat{h})}{\text{Prob}(h)}$
Female age 18–34	25.18	58.06	2.30
Female age >55	28.65	76.15	2.66
Male age 12–17	11.68	31.91	2.73
Male age 18–34	22.99	60.78	2.64
Male or female age 2–11	24.82	84.21	3.39

Note: The age-gender combinations included here are illustrative, and therefore do not include every segment. The ages shown can be characterized as *children* (2–11), *teenagers* (12–17), *young adults* (18–34), and *older adults* (>55).

among binary classifiers to determine the strength of association between accuracy and each individual diversity measure [21]. They used an *oracle* representation, which does not distinguish between types of outcome errors. It assumes misclassification costs are identical. However, in a targeting advertising environment, misclassification costs are not equal because errors in accuracy are much more costly than errors in sensitivity. Predicting that an observation is a member of the target set is a business commitment to an advertiser and a financial commitment from the advertiser, and therefore carries a concrete dollar cost when in error. Ads have to be sent to x/a_c households in order to have an expected yield of x exposures to the target audience, so the higher a_c is, the lower the cost of the advertising campaign. As discussed below, some of Kuncheva and Whitaker's conclusions extend to the unequal misclassification case, but we found that the relationship between the diversity measures and performance is a function of both population and individual classifiers' characteristics.

Within the context of ensemble binary classification with simple majority voting, we address the following research questions:

- Which numerical diversity measures appear to be statistically significant predictors of the performance of an ensemble, where the predictive model is of arbitrary form and may involve multiple predictors?
- In extending Kuncheva and Whitaker's study, do the models for the direct representation differ from those estimated using an oracle representation?
- Do variations on the diversity measures used by Kuncheva and Whitaker, or additional diversity measures, improve ensemble predictability?
- How do specific characteristics of the domain, such as percentage of the target group in the population and performance of individual models, impact the ability of diversity measures to predict ensemble performance?
- Would combinations of diversity measures make better predictors of ensemble performance than individual measures?

The rest of the paper is organized as follows. In Section 2 we define concepts and terms related to classifier performance and representation. In Section 3 we discuss the diversity measures explored in this research, while in Section 4 we describe our research approach, including the individual experiments and their design parameters. Section 5 describes the results from the experiments within and across representations. Section 6 then concludes the paper with a discussion of the major findings and their implications.

2. Performance and representation

Ensemble classification is based on the assumption that a more diverse group of classifiers may perform better than a less diverse one. As mentioned in Section 1, we use accuracy and sensitivity to measure performance. While the definition of the term tends to vary (see [14], pp. 98–99 and p. 182), we define *accuracy* ($\mathbf{a}_c$) as measuring the percentage of observations predicted by a model to be members of a target group (i.e., predicted 1s) which actually are members of the group (i.e., actual 1s). Alternatively stated, it is the probability that an observation is a 1 given that it is predicted as a 1; i.e., $P(actual = 1|prediction = 1)$. *Sensitivity* in turn is the conditional probability of predicting that an observation is a 1 given that it is a 1; i.e., $P(prediction = 1|actual = 1)$. The complementary conditional probability, $P(prediction = 0|actual = 1) = (1 - sensitivity)$ is an opportunity cost, measuring the proportion of the available inventory (i.e., members of a particular demographic group) not recognized as such by the model.

To illustrate the issue of performance in the context of both symmetric and asymmetric costs, let $\mathbf{p}$ denote the proportion of vectors in S that are actually members of the group labeled 1. For any classifier or ensemble, the direct representation retains the assigned and true group labels for each item in S, which can be cross-classified into a *confusion matrix* of the form shown in Table 2. Using standard ROC (receiver operating characteristic) curve terminology, the specificity $\mathbf{sp}$ and sensitivity $\mathbf{se}$ are, respectively, the proportion of true negatives and the proportion of true positives achieved by a classifier or ensemble. Specificity and sensitivity are calculated as follows:

$$\mathbf{sp} = \mathbf{a}/(\mathbf{a} + \mathbf{c})$$

Table 2
Confusion matrix showing classifier performance in the context of both symmetric and asymmetric costs, where **p** is the proportion of vectors in S that are actually members of the group labeled 1, specificity (**sp**) is the proportion of true negatives achieved by a classifier or ensemble, and sensitivity (**se**) is the proportion of true positives

		Predicted		
		0	1	
Actual	0	$(1-\mathbf{p})\mathbf{sp}$	$(1-\mathbf{p})(1-\mathbf{sp})$	$(1-\mathbf{p})$
	1	$\mathbf{p}(1-\mathbf{se})$	$\mathbf{p}(\mathbf{se})$	$\mathbf{p}$
				1

$\mathbf{se} = \mathbf{d}/(\mathbf{b} + \mathbf{d})$

where **a** is the number of actual 0s predicted as 0, **b** is the number of actual 1s predicted as 0, **c** is the number of actual 0s predicted as 1, and **d** is the number of actual 1s predicted as 1 ([14], pp. 131–140).

For reasons discussed later, we measure the quality of the output of a classifier or ensemble using sensitivity and the proportion of correct predictions of membership in the group labeled 1, the *accuracy using the direct representation*, $\mathbf{a}_c$. Because $\mathbf{a}_c = \mathbf{p}(\mathbf{se})/[(1-\mathbf{p})(1-\mathbf{sp}) + \mathbf{p}(\mathbf{se})]$, accuracy using the direct representation is a function of both specificity and sensitivity (as derived from [14]). By contrast, the oracle representation retains only the correctness or incorrectness of the classifier's or ensemble's labeling, not the assigned or true group membership. Thus, the *accuracy using the oracle representation*, $\mathbf{a_o}$, is the sum of the principal diagonal entries in the confusion matrix, that is, $\mathbf{a_o} = (1-\mathbf{p})\mathbf{sp} + \mathbf{p}(\mathbf{se})$.

Identifying and generalizing from the conditions impacting ensemble performance requires a high level of control over the situational factors comprising the problem domain. Because the results obtained from analysis of real-world data cannot be sufficiently controlled, we have pursued an experimental approach—starting with the simulated output of multiple classification methods, continuing with the generation of consensus classifications from those simulated results, and ending with an understanding of the factors impacting combination strategies. Simulated data sets that are not linked per se to any particular domain provide a much higher level of control over the situational parameters that might impact combination strategies, while avoiding many of the idiosyncrasies of real-world data and data mining methods. The simulated data reflect the characteristics and constraints of the television viewing data upon which they are based, and therefore comprise a realistic—even though generated—representation of the actual viewing data.

3. Diversity measures

We used 22 diversity measures for each ensemble, comprised of 15 pairwise and 7 non-pairwise measures. The 15 pairwise measures are comprised of three values—the average across the members of the ensemble, as well as the maximum, and minimum—for each of 5 pairwise measures. The formulas used to calculate the measures are shown below (*See Kuncheva and Whitaker* [21] *for a complete description of measures* (1)–(4). *We calculated those measures in the same manner and use the same notation below. Note that Kuncheva and Whitaker restrict their calculations to the average pairwise values.*)

Table 3
2×2 matrix showing the relationship between a pair of classifiers, where N^{xy} is the number of observations for which classifier 1 predicts x (0,1) and classifier 2 predicts y (0, 1)

		Classifier 2	
		1	0
Classifier 1	1	N^{11}	N^{10}
	0	N^{01}	N^{00}

The total number of observations $N = N^{11} + N^{10} + N^{01} + N^{00}$ (*adapted from* [21]).

1. Yule's Q-statistic (Q) [36] for two classifiers, D_i and D_k, is

   $$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}$$

   where N is the total number of predictions, and N^{xy} is the number of predictions for each intersection of D_i and D_k—as shown in Table 3. The average over all pairs of L classifiers is:

   $$Q_{\text{av}} = \frac{2}{L(L-2)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k}$$

2. The correlation coefficient (ρ) for two binary classifier outputs y_i and y_k is

   $$\rho_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11} + N^{10})(N^{01} + N^{00})(N^{11} + N^{00})(N^{10} + N^{00})}}$$

3. The disagreement measure (Dis) [33] is

   $$\text{Dis}_{i,k} = \frac{N^{01} + N^{10}}{N^{11}N^{10} + N^{01}N^{00}}$$

4. The double-fault measure (DF) [13] is

   $$\text{DF}_{i,k} = \frac{N^{00}}{N^{11} + N^{10} + N^{01} + N^{00}}$$

5. The general chi-square value (χ^2) for deviation from independence ([6, pp. 204–215]) is defined as follows. Let $R_i = \sum_j N_{i,j}$, $C_j = \sum_i N_{i,j}$ and $N = \sum_i \sum_j N_{i,j}$ where R_i is the sum of the ith row of the confusion matrix, C_j is the sum of the jth column, and N is the total number of observations.
   Then the observed chi-square value is the sum of the squared differences of the observed minus the expected cell frequencies divided by the expected cell frequencies. That is,

   $$\chi^2 = \frac{\sum_i \sum_j (N_{i,j} - E_i)}{E_i}, \quad \text{where } E_i = \frac{R_i C_i}{N}$$

The remaining seven non-pairwise diversity measures, derived from all the models in an ensemble, are described below (note that Kuncheva and Whitaker used the non-pairwise measures (1)–(6); again, we calculated the measures identically).

1. The Kohavi–Wolpert variance (KW) [17] is

$$\mathrm{KW} = \frac{1}{NL}\sum_{j=1}^{N} l(z_j)(L - l(z_j))$$

where z_j is the jth element of a data set $Z = \{z_1, z_2, \ldots, z_N\}$

2. The interrater agreement (κ) [11] can be calculated from the KW variance, first by calculating the average individual classification accuracy ($\overline{p}$):

$$\overline{p} = \frac{1}{NL}\sum_{j=1}^{N}\sum_{i=1}^{L} y_{j,i}$$

κ is then calculated from KW, $\overline{p}$ and L:

$$\kappa = 1 - \frac{L}{(L-1)\overline{p}(1-\overline{p})}\mathrm{KW}$$

3. The entropy measure (E) varies between 0 and 1, where 0 indicates no diversity among classifiers, and 1 indicates maximum diversity:

$$E = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{(L - [L/2])}\min\{l(z_j), L - l(z_j)\}$$

4. The measure of difficulty (θ) [15] is a measure of the variance of X:

$$\theta_a = \mathrm{Var}(X_a)$$

where X is a discrete random variable with values $\{\frac{0}{L}, \frac{1}{L}, \ldots, 1\}$ indicating the proportion of classifiers in the ensemble correctly classifying a randomly drawn input (x) from the problem distribution (see also [21]).

5. The generalized diversity (GD) measure [27] is calculated from the probability of failure of one or both of two classifiers drawn randomly from an ensemble, where $p(1) =$ probability that one classifier will fail and $p(2) =$ probability that both classifiers will fail. $p(1), p(2)$, and the generalized diversity measure then are calculated as follows:

$$p(1) = \sum_{i=1}^{L}\frac{i}{L}p_i \quad \text{and} \quad p(2) = \sum_{i=1}^{L}\frac{i(i-1)}{L(L-1)}p_i$$

$$\mathrm{GD} = 1 - \frac{p(2)}{p(1)}$$

6. The coincident failure diversity (CFD) [27] is a variation on the GD measure (as noted in [21])

$$\mathrm{CFD} = \begin{cases} 0, & p_0 = 1.0 \\ \frac{1}{1-p_0}\sum_{i=1}^{L}\frac{L-i}{L-1}p_i, & p_0 < 1.0 \end{cases}$$

7. We then use the same general chi-square test for independence used in the pairwise analysis.

4. Research approach

Formally, we consider a binary classification task in which the two classes are labeled 0 and 1, and a set $M = \{M_1, M_2, \ldots, M_c\}$ of classifier methods that assign label $M_i(x) \in \{0, 1\}$ to N vectors $x_1, x_2, \ldots, x_N$ (where $\mathbf{x}$ is the vector of values of the independent variables and N is the number of items classified—as indicated below). Let l be a subset of distinct member of the set of indices $\{1, 2, \ldots, c\}$, where the cardinality of l is L. We assume that L is odd. Let the labeling assigned by the ensemble involving the indices in l, $E(l)$ be given by $E(l) = \lfloor 1/2 + (\Sigma_{i \in l} M_i / l \rfloor$, where $\lfloor\ \rfloor$ is the integer floor, that is, we use simple majority vote to assign the ensemble's label. Simple majority vote in our domain is feasible across all ensemble sizes because of the voting results are binary (i.e., restricted to either 0 or 1), and because ensembles are assumed at present to contain an odd number of members. Note that we deduce the ensemble classification only from the labels $M_i(x)$ assigned by the methods, not from the values of the underlying vectors.

4.1. Design parameters

The design parameters and values used include:

- *Number of items classified* (N): the number of individuals to be classified by each method. The results should not be sensitive to the value of N, except that the larger the value of N, the more possible values in [0,1] that can be achieved by the ensemble. We used $N = 1000$.
- *Percent in the population* ($\mathbf{p}$): the fraction of the general population that belongs to the target group. We used 10%, 20%, 30%, 40%, 50% and 60%, because in a television viewing application, those are typical values for target audiences.
- *Total number of models* (c) from which an ensemble can be created. This and L determine the number of replications for each set of parameters, because we generate all ${}_c\mathrm{C}_L$ ensembles of L items chosen from the c possible. We used $c = 10$.
- *Number of models* (L) *that participate in the ensemble*. We set $L = 3$, because it avoids tie votes while also keeping the number of replications manageable. Thus there are ${}_{10}\mathrm{C}_3 = 120$ replications for each set of parameters.
- *Designed accuracy* ($\mathbf{a}_\mathrm{c}$) *and sensitivity of each model* ($\mathbf{se}$): each model label vector has the same range of values for $\mathbf{a}_\mathrm{c}$ and $\mathbf{se}$.
- *Voting policy*: Currently, the consensus prediction of an ensemble is achieved through a simple majority (democratic) voting procedure, which is facilitated by having an odd number of voting models in each ensemble. While democratic voting is an effective and popular approach [23,24], other more complex voting schemes are potentially effective and could be employed based on a variety of criteria [16]. (We discuss later the potential for extending this research

Table 4
Accuracy ($\mathbf{a}_c$), sensitivity (**se**) and percent-in-population (**p**) parameter values across 60 experiments

Experiment	($\mathbf{a}_c$) (%)	**se** (%)	**p** (%)
1–6	40	20	10–60
7–12	40	40	10–60
13–18	60	20	10–60
19–24	60	40	10–60
25–30	60	60	10–60
31–36	60	80	10–60
37–42	80	20	10–60
43–48	80	40	10–60
49–54	80	60	10–60
55–60	80	80	10–60

through manipulations of voting policy and randomized outcomes.)

4.2. Experiments

We designed a series of 60 experiments that were defined as a specific combination of parameter values that varied across the experiments. Table 4 indicates how the parameters were distributed to each set of experiments. For example, in the first set of six experiments, designed accuracy and sensitivity were fixed at 40% and 20% respectively, while percent in the population varied from 10% to 60%. In the second set, accuracy remained at 40%, sensitivity was changed to 40%, and again percent in the population varied from 10% to 60%. The remaining sets of experiments were designed similarly.

We should note that these values are not all-encompassing, but rather represent realistic constraints on the values of the parameters based on the characteristics of our problem domain. In advertising, the proportion of the target group relative to the population at-large generally is relatively small, which led us to restrict the percent in the population parameter to less than 60% (resulting in correspondingly high values for specificity). Furthermore, certain parameter value combinations—for example, accuracy $[P(actual = 1|prediction = 1)]$ of 40%, sensitivity $[P(prediction = 1|actual = 1)]$ of 60%, and percent in population of 60% are impossible because those values require that $P(actual = 0$ and $prediction = 0)$ be less than zero; −14% in the case cited.

The data initialization phase of the experiment then generated performance and diversity measures in the context of the situational parameters defined for each of the experiments described above. This phase proceeded in the following sequence:

1. *generation of the actual group assignment vector*, against which an ensemble's performance is assessed. In a real data set, this vector would indicate whether or not an individual belongs to the target class (i.e., either 1 or 0). In the experimental approach, the vector is populated with 1s and 0s depending on the value of the percent in the population (**p**) parameter. For example, if the value of **p** is 20 (percent) and the number of individuals (N) is 1000, then 200 of the values in the vector would be '1', with the rest being '0';
2. *generation of c model prediction vectors*, which represent the predictions of c different hypothetical models. The number and content of the vectors are based on the value of c defined in the database, along with the accuracy and sensitivity rates specified for each of the prediction vectors;
3. *assembly of subsets of the* $(c = 10)$ *prediction vectors into ensembles of* $(L = 3)$ *voting vectors;*
4. *generation of* $({}_cC_L = 120)$ *ensemble predictions* for each individual by tallying the votes of each of the participating voting vectors within each ensemble;
5. *calculation of the accuracy and sensitivity rates of each ensemble*, which determines the likelihood of predicting;
6. *determination of the ensemble's performance* by comparing it to the rates of the individual vectors/models within the ensemble;
7. *calculation of pairwise and non-pairwise diversity measures* that relate pairs and groups of the models in each ensemble across all the observations and for each of the two groups separately, for each of two alternative representations of the underlying classification problem.

The performance and diversity measures then were fed into neural network models (SPSS/Clementine) to search for patterns showing which of the diversity measures are most indicative of ensemble performance—for both the oracle and direct representations. The neural network used an exhaustive prune search to identify the significant diversity measures. Exhaustive pruning is similar to stepwise regression, in that only the diversity measures (i.e., independent variables) considered significant in predicting performance are retained in the network when it is trained. Generalizability of the model was enhanced by withholding half of the sample data when training the network, and then using the other half to control overfitting. The random seed was held constant across all models at a value of 12,345. We essentially deferred the settings of the other model parameters to Clementine, so that it would construct the best model given the constraints imposed by exhaustive pruning. Thus, parameters such as the number of nodes and hidden layers were determined by the model building routine in Clementine.

5. Results

Results from the experiments, which can be described across and within representations, are both consistent with, and different from, those of Kuncheva and Whitaker. Comparing across representations, Yule's Q statistic (Q_{avg}) is the best indicator of ensemble performance in the direct representation—where it is tied with the coincident failure diversity (CFD) measure—but not in the oracle representation, where CFD is the best indicator of ensemble performance (see Table 5 Panels A and B, *Total* column). This finding is notable in that it supports Kuncheva and Whitaker's finding that the Q-statistic is the best performance indicator, but only in one of the representations.

It also is notable that within each data representation, the association between diversity measures and ensemble performance is contingent on the values of three manipulated experimental parameters: percent in

Table 5
Number of diversity measure appearances as a function of $\mathbf{p}$ for the oracle representation (Panel A) and direct representation (Panel B)

Diversity measure	Number of appearances						
	$\mathbf{p} = 0.1$	$\mathbf{p} = 0.2$	$\mathbf{p} = 0.3$	$\mathbf{p} = 0.4$	$\mathbf{p} = 0.5$	$\mathbf{p} = 0.6$	Total
Panel A							
Q_{avg}	7	3	3	2	5	6	26
Q_{max}	5	4	6	3	3	1	22
Q_{min}	4	2	4	3	3	3	19
ρ_{avg}	6	0	0	3	2	2	13
ρ_{max}	0	2	4	3	3	3	15
ρ_{min}	4	0	1	0	2	1	8
D_{avg}	1	2	5	6	7	5	26
D_{max}	4	5	5	2	5	4	25
D_{min}	3	1	5	3	4	3	19
$\mathrm{DF}_{\mathrm{avg}}$	2	0	3	0	2	2	9
$\mathrm{DF}_{\mathrm{max}}$	1	0	2	0	2	2	7
$\mathrm{DF}_{\mathrm{min}}$	1	0	3	2	2	2	10
χ^2_{avg}	5	3	5	3	2	2	20
χ^2_{max}	3	2	3	3	4	3	18
χ^2_{min}	1	2	1	2	2	1	9
E	4	5	6	4	7	4	30
KW	1	2	5	4	7	4	23
κ	4	0	5	4	4	5	22
θ	5	1	3	3	5	3	20
GD	6	6	6	3	7	4	32
CFD	10	9	9	10	9	9	56
χ^2	9	6	7	4	4	5	35
Max possible	10	10	10	10	10	10	60
Panel B							
Q_{avg}	6	5	7	5	9	2	34
Q_{max}	5	1	7	5	5	2	25
Q_{min}	5	1	4	2	5	2	19
ρ_{avg}	6	2	5	4	5	1	23
ρ_{max}	3	1	6	3	7	0	20
ρ_{min}	3	1	4	1	3	1	13
D_{avg}	3	2	5	3	6	1	20
D_{max}	1	0	3	3	4	0	11
D_{min}	3	0	5	3	4	1	16
$\mathrm{DF}_{\mathrm{avg}}$	4	0	2	0	2	0	8
$\mathrm{DF}_{\mathrm{max}}$	3	0	4	1	2	0	10
$\mathrm{DF}_{\mathrm{min}}$	3	0	2	1	2	0	8
χ^2_{avg}	7	4	4	3	3	0	21
χ^2_{max}	3	0	5	1	5	0	14
χ^2_{min}	3	0	5	2	5	0	15
E	3	2	5	3	3	2	18
KW	4	0	4	2	6	1	17
κ	8	4	4	3	4	3	26
θ	6	3	5	3	4	2	23
GD	1	0	2	1	1	1	6
CFD	5	4	7	6	6	6	34
χ^2	8	3	6	0	2	3	22
Max possible	10	10	10	10	10	10	60

the population (**p**), designed direct accuracy ($\mathbf{a}_c$), and designed sensitivity (**se**). The results for the two leading measures in both representations, Q_{avg} and CFD, are charted in Figs. 1–3 (for **p**, $\mathbf{a}_c$ and **se**, respectively). The X-axis of each graph indicates the manipulated levels of the parameter, while the Y-axis shows the percentage of times that Q_{avg} and CFD appeared in a final neural network exhaustive prune model. The Y-axis therefore indicates how well a particular measure predicts ensemble performance at each parameter level. As shown in Fig. 1, in the direct representation, Q_{avg} is a better indicator of performance than CFD at certain values of **p** (i.e., 0.1, 0.2 and 0.5), while CFD is better than Q_{avg} at other values (i.e., 0.4 and 0.6). The same patterns are evident when the parameter-of-interest is $\mathbf{a}_c$ and **se**—i.e., Q_{avg} is better at some values while CFD is better at others. In the oracle representation, by contrast, CFD is the best indicator of performance for all values of all three parameters. While Figs. 1–3 show that different parameter values appear to impact the utility of CFD in predicting ensemble performance, the impact is small relative to the fluctuations shown by Q_{avg}. That is, Q_{avg} association with performance is much more volatile within each representation, and thus much more dependent on the specific value of a parameter.

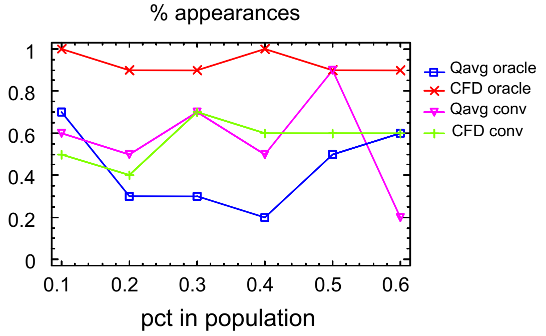

Fig. 1. Percentage of appearances in a neural network model for Q_{avg} and CFD in the oracle and direct representations, as percent-in-population (**p**) increases from 0.1 to 0.6.

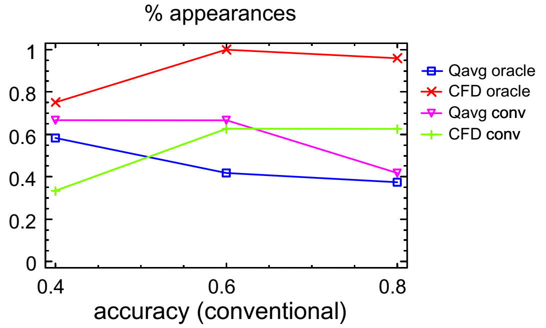

Fig. 2. Percentage of appearances in a neural network model for Q_{avg} and CFD in the oracle and direct representations, as designed direct accuracy ($\mathbf{a}_c$) increases from 0.4 to 0.8.

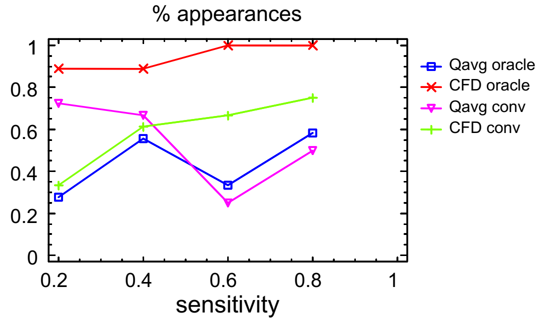

Fig. 3. Percentage of appearances in a neural network model for Q_{avg} and CFD in the oracle and direct representations, as designed sensitivity (**se**) increases from 0.2 to 0.8.

Tables 5–7 present more detailed results which include all of the diversity measures. As a group, the tables show the influence of each diversity measure, in both representations, as **p** (Table 5), $\mathbf{a}_c$ (Table 6) and **se** (Table 7) vary across the range of values. Each table shows (1) the number of times each measure appeared in a final neural network model for each value of the parameter of interest, and (2) the total number of times the measure appeared across all models (the latter is identical for each representation across all three tables). Across all parameters, CFD is the most significant in 56 of the 60 experiments in the oracle representation (Table 5, Panel A)—followed distantly by the other measures. In the direct representation (Table 5, Panel B), Q_{avg} and CFD essentially are tied—each showing significance in 34 experiments.

Within each of the individual parameters (**p**, $\mathbf{a}_c$ and **se**), the relationship between measure and performance varies substantially across the two representations. The results relative to each of the parameters are discussed below.

5.1. Percent in population

For the percent-in-population (**p**) parameter, CFD is consistently most related to performance in the oracle representation (Table 5, Panel A) across all levels of **p**, showing significance in at least 9 of 10 experiments at each level. However, in the direct representation (Table 5, Panel B), different diversity measures are better indicators of performance at different levels of **p**. For example, although Q_{avg} and CFD are most closely related to performance overall, Q_{avg} is the single best indicator only when $\mathbf{p} = 0.2$, 0.3 and 0.5, while CFD is the best indicator when $\mathbf{p} = 0.3$, 0.4 and 0.6. Notably, when $\mathbf{p} = 0.1$, interrater agreement (κ) and χ^2 (non-pairwise) become the best indicators.

Table 6
Number of diversity measure appearances as a function of ($\mathbf{a}_c$) for the oracle representation (Panel A) and direct representation (Panel B)

Diversity measure	Number of appearances			
	$\mathbf{a}_c = 40\%$	$\mathbf{a}_c = 60\%$	$\mathbf{a}_c = 80\%$	Total
Panel A				
Q_{avg}	7	10	9	26
Q_{max}	6	6	10	22
Q_{min}	2	7	10	19
ρ_{avg}	4	3	6	13
ρ_{max}	3	4	8	15
ρ_{min}	2	1	5	8
D_{avg}	5	10	11	26
D_{max}	3	10	12	25
D_{min}	4	8	7	19
DF_{avg}	1	4	4	9
DF_{max}	0	5	2	7
DF_{min}	2	5	3	10
χ^2_{avg}	3	10	7	20
χ^2_{max}	3	7	8	18
χ^2_{min}	0	4	5	9
E	6	13	11	30
KW	5	11	7	23
κ	6	7	9	22
θ	6	8	6	20
GD	5	14	13	32
CFD	9	24	23	56
χ^2	7	14	14	35
Max possible	12	24	24	60
Panel B				
Q_{avg}	8	16	10	34
Q_{max}	5	11	9	25
Q_{min}	5	10	4	19
ρ_{avg}	2	12	9	23
ρ_{max}	3	6	11	20
ρ_{min}	1	10	2	13
D_{avg}	3	10	7	20
D_{max}	2	5	4	11
D_{min}	4	7	5	16
DF_{avg}	1	4	3	8
DF_{max}	0	7	3	10
DF_{min}	0	4	4	8
χ^2_{avg}	2	9	10	21
χ^2_{max}	1	7	6	14
χ^2_{min}	2	7	6	15
E	2	9	7	18
KW	2	8	7	17
κ	4	9	13	26
θ	2	9	12	23
GD	0	4	2	6
CFD	4	15	15	34
$\chi 2$	2	9	11	22
Max possible	12	24	24	60

5.2. Accuracy

Similarly, varying the designed accuracy parameter ($\mathbf{a}_c$) has a differential impact on the diversity measures indications of performance. While CFD remains the best indicator across all three levels of $\mathbf{a}_c$ in the oracle representation (Table 6, Panel A), the number of experiments in which it is significant is not consistent—unlike the results described above for different levels of **p**. When $\mathbf{a}_c = 60\%$ and 80%, CFD is by far the leading measure. CFD shows significance in 24 and 23 experiments (out of 24), respectively, while the next best indicator is significant only in 14 out of 24 experiments. However, when $\mathbf{a}_c = 40\%$, CFD is significant in 9 out of

Table 7
Number of diversity measure appearances as a function of **se** for the oracle representation (Panel A) and direct representation (Panel B)

Diversity measure	Number of appearances				
	se = 20%	**se** = 40%	**se** = 60%	**se** = 80%	Total
Panel A					
Q_{avg}	5	10	4	7	26
Q_{max}	5	8	3	6	22
Q_{min}	2	7	4	6	19
ρ_{avg}	4	3	3	3	13
ρ_{max}	4	3	4	4	15
ρ_{min}	2	3	1	2	8
D_{avg}	8	7	5	6	26
D_{max}	5	7	7	6	25
D_{min}	4	6	5	4	19
DF_{avg}	1	1	3	4	9
DF_{max}	1	0	3	3	7
DF_{min}	2	1	4	3	10
χ^2_{avg}	10	6	1	3	20
χ^2_{max}	4	7	3	4	18
χ^2_{min}	2	2	3	2	9
E	7	8	9	6	30
KW	6	6	5	6	23
κ	6	6	4	6	22
θ	6	6	4	4	20
GD	6	9	10	7	32
CFD	16	16	12	12	56
$\chi 2$	12	9	8	6	35
Max possible	18	18	12	12	60
Panel B					
Q_{avg}	13	12	3	6	34
Q_{max}	11	7	3	4	25
Q_{min}	9	5	2	3	19
ρ_{avg}	7	3	5	8	23
ρ_{max}	9	4	2	5	20
ρ_{min}	3	3	3	4	13
D_{avg}	4	4	4	8	20
D_{max}	3	3	2	3	11
D_{min}	6	1	1	8	16
DF_{avg}	4	1	0	3	8
DF_{max}	3	2	2	3	10
DF_{min}	3	0	1	4	8
χ^2_{avg}	2	4	6	9	21
χ^2_{max}	4	2	3	5	14
χ^2_{min}	4	2	4	5	15
E	2	4	4	8	18
KW	3	4	3	7	17
κ	7	7	3	9	26
θ	7	4	4	8	23
GD	2	0	3	1	6
CFD	6	11	8	9	34
χ^2	4	6	5	7	22
Max possible	18	18	12	12	60

12 experiments, while χ^2 (non-pairwise) and Q_{avg} are significant in 7 experiments. Thus, CFD is a better indicator *relative to the other measures* at higher levels of $\mathbf{a}_c$, and less so at lower levels of $\mathbf{a}_c$. At the same time, Q_{avg} becomes a relatively better indicator when $\mathbf{a}_c$ is a lower value (i.e., 40%). Although Q_{avg} is well behind CFD as an overall indicator (appearing in 26 NN models compared to 56 for CFD), it is tied for 2nd when $\mathbf{a}_c = 40\%$, appearing in only two fewer models than CFD at that level of $\mathbf{a}_c$.

5.3. Sensitivity

The patterns for variations in the sensitivity parameter (**se**) are similar to those shown for accuracy. In the oracle representation, CFD is the leading indicator of performance across values of sensitivity, although again it tends to vary. For **se** = 40% and 80%, CFD is significant in substantially more models than the second place measure (16 vs. 10 and 12 vs. 7). For **se** = 20% and 60%, the gap between CFD and its nearest competitor shrinks somewhat (16 vs. 12 and 12 vs. 10). In the direct representation, Q_{avg} is the leading measure at the lower levels of **se**, 20% and 40%, while it trails CFD at the higher levels of 60% and 80%.

The specific impact of **se** on the χ^2_{avg} measure is notable, since χ^2_{avg} becomes a substantially better indicator of performance as the value of **se** increases. At **se** = 20%, it appears in only 2 out of 18 measures (11%), and is essentially the worst indicator of ensemble performance. As **se** increases to 80%, the number of models within which χ^2_{avg} is significant rises to 9 out of 12 (75%), where it becomes the *best* indicator of ensemble performance (although tied with CFD and κ).

6. Conclusions

The major findings of this research show that the ability of diversity measures to predict ensemble performance varies depending on (1) the type of data representation used, either direct or oracle, and (2) the values of specific situational parameters within each of the data representations. These are significant because, in a decision making task, they suggest the use of a contingent approach in selecting diversity measures for performance evaluation, in several respects. First, the choice of diversity measures becomes contingent on data representation, and by extension, on the impact of misclassification costs. In the context of asymmetric misclassification costs—i.e., within the direct representation—Yule's Q-statistic is a relatively good indicator of performance. But in the oracle representation, wherein misclassification costs are treated identically, the Q-statistic appears less effective than several other measures, particularly the coincident diversity (CFD) measure.

Second, as noted, the choice of diversity measures also is contingent on variations in situational parameters, particularly within the direct representation. That is, in a domain characterized by asymmetric misclassification costs, Yule's Q-statistic is a preferred measure of performance at certain parameter levels, but not at others. By contrast, in a domain characterized by symmetric misclassification costs, CFD is the best choice across all levels of each parameter—although other diversity measures become more competitive with CFD at certain parameter levels.

Furthermore, the imposed data representation affects the relevancy of some parameters. For example, the percent-in-population (**p**) parameter is relevant only in the direct representation because it directly impacts the calculated values of accuracy and sensitivity. In this representation, because both accuracy and sensitivity are determined based on whether an actual observation is a member of the target group, knowing the percentage of members in the population is required. By contrast, because the oracle representation records predictions as either correct or incorrect, without regard to the value of the prediction itself, knowing the percentage of members in the population is not required.

By exploring the impact of data representation and situational factors, this research has sought answers to the general 'open question' of diversity measure utility posed by Kuncheva and Whitaker. But a number of questions remain. While this study contributes to a general understanding of how diversity measures are impacted by situational factors, we do not yet understand, for example, why Yule's Q and CFD are better relative indicators of performance, or why their performance varies across parameter values. More work across the research community is required to discover what other factors might be important, and to learn how and why the various elements of the classification problem impact the choice of diversity measures. Several of the parameters held constant in this study are amenable to further exploration. They include the total number of models generated (c), the number of participating models (L), the randomized outcome of designed accuracy and sensitivity, and the voting policy—which can be extended beyond the simple democratic process used in this study to include more complex, weighted approaches. Because the ability of diversity measures to reflect ensemble performance is contingency-based, exploration of these additional factors will help in building a consensus understanding of how and why diversity measures—and perhaps even diversity itself—are important to ensemble performance.

References

[1] K.M. Ali, M.J. Pazzani, Error reduction through learning multiple descriptions, Machine Learning 24 (3) (1996) 173–202.

[2] R.E. Banfield, L.O. Hall, K.W. Bowyer, W.P. Kegelmeyer. A new ensemble diversity measure applied to thinning ensembles, in: 4th International Workshop on Multiple Classifier Systems, Guildford, UK, 2003.

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

[4] P.K. Chan, S.J. Stolfo, On the accuracy of meta-learning for scalable data mining, Journal of Intelligent Information Systems 8 (1997) 5–28.

[5] L. Chunovic, Change ahead for channels, Electronic Media 20 (31) (2001) 1.

[6] W.J. Conover, Practical Nonparametric Statistics, third ed., John Wiley & Sons, New York, 1999.

[7] P. Cunningham, J. Carney, Diversity versus quality in classification ensembles based on feature selection, in: 11th European Conference on Machine Learning, Barcelona, Spain, 2000.

[8] T.H. Davenport, Saving IT's soul: human-centered information management, Harvard Business Review (March–April 1994) 119–131.

[9] T.G. Dietterich, Machine learning research: four current directions, AI Magazine 18 (4) (1997) 97–136.

[10] T.G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning 40 (2) (2000) 139–157.

[11] J. Fleiss, Statistical Methods for Rates and Proportions, John Wiley & Sons, New York, 1981.

[12] Y. Freund, R. Schapire, Experiments with a new boosting algorithm, in: 13th International Conference on Machine Learning, Bari, Italy, 1996.

[13] G. Giacinto, F. Roli, Design of effective neural network ensembles for image classification purposes, Image Vision and Computing Journal 19 (9–10) (2001) 697–705.

[14] D.J. Hand, Construction and Assessment of Classification Rules, John Wiley & Sons, Chichester, 1997.

[15] L. Hansen, P. Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (10) (1990) 993–1001.

[16] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.

[17] R. Kohavi, D. Wolpert, Bias plus variance decomposition for zero-one loss functions, in: Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kaufmann, Bari, Italy, 1996, pp. 275–283.

[18] A. Krogh, J. Vedelsby, Neural network ensembles cross validation and active learning, in: G. Tesauro, D.S. Touretzky, T.K. Leen (Eds.), Advances in Neural Information Processing Systems, MIT Press, Cambridge, Mass, 1995.

[19] L.I. Kuncheva, J. Beezdek, R.P.W. Duin, Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition 34 (2) (2001) 299–314.

[20] L.I. Kuncheva, M. Skurichina, R.P.W. Duin, An experimental study on diversity for bagging and boosting with linear classifiers, Information Fusion 3 (2) (2002) 245–258.

[21] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning 51 (2) (2003) 181–207.

[22] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W. Duin, Limits on the majority vote accuracy in classifier fusion, Pattern Analysis Applications 6 (2003) 22–31.

[23] L. Lam, C.Y. Suen, Optimal combinations of pattern classifiers, Pattern Recognition Letters 16 (1995) 945–954.

[24] L. Lam, C.Y. Suen, Application of majority voting to pattern recognition: an analysis of its behavior and performance, IEEE Transactions on Systems, Man and Cybernetics 27 (5) (1997) 553–568.

[25] B. Littlewood, D. Miller, Conceptual modeling of coincident failures in multiversion software, IEEE Transactions on Software Engineering 15 (12) (1989) 1596–1614.

[26] D. Opitz, R. Maclin, Popular ensemble methods: an empirical study, Journal of Artificial Intelligence Research 11 (1999) 169–198.

[27] D. Partridge, W. Krzanowski, Software diversity: practical statistics for its measurement and exploitation, Information and Software Technology 39 (1997) 707–712.

[28] E. Pekalska, R.P.W. Duin, Dissimilarity representations allow for building good classifiers, Pattern Recognition Letters 23 (2002) 943–956.

[29] B.E. Rosen, Ensemble learning using decorrelated neural networks, Connection Science 8 (3) (1996) 373–384.

[30] D. Ruta, B. Gabrys, New measure of classifier dependency in multiple classifier systems, in: 3rd International Workshop on Multiple Classifier Systems, Cagliari, Sardinia, Italy, 2002, pp. 127–136.

[31] R. Schapire, The strength of weak learnability, Machine Learning 5 (2) (1990) 197–227.

[32] C.A. Shipp, L.I. Kuncheva, Relationships between combination methods and measures of diversity in combining classifiers, Information Fusion 3 (2) (2002) 135–148.

[33] D. Skalak, The sources of increased accuracy for two proposed boosting algorithms, in: Proceedings of the AAAI-96 Integrating Multiple Learned Models Workshop Portland, AAAI Press, Ore., 1996, pp. 120–125.

[34] W.E. Spangler, M. Gal-Or, J.H. May, Using data mining to profile television viewers in the digital TV era, Communications of the ACM 46 (4) (2003) 66–72.

[35] C.J. Whitaker, L.I. Kuncheva, Examining the relationship between majority vote accuracy and diversity in bagging and boosting, School of Informatics, University of Wales, Bangor, 2003.

[36] G.U. Yule, On the association of attributes in statistics, Philosophical Transactions of the Royal Society of London, Series A 194 (1900) 257–319.