Research Article

Journal of Algorithms & Computational Technology

The effectiveness of using diversity to select multiple classifier systems with varying classification thresholds

Journal of Algorithms & Computational Technology
2018, Vol. 12(3) 187–199

DOI: 10.1177/1748301818761132
journals.sagepub.com/home/act
SAGE

Harris K Butler IV, Mark A Friend, Kenneth W Bauer Jr and Trevor J Bihl

Abstract

In classification applications, the goal of fusion techniques is to exploit complementary approaches and merge the information provided by these methods to provide a solution superior than any single method. Associated with choosing a methodology to fuse pattern recognition algorithms is the choice of algorithm or algorithms to fuse. Historically, classifier ensemble accuracy has been used to select which pattern recognition algorithms are included in a multiple classifier system. More recently, research has focused on creating and evaluating diversity metrics to more effectively select ensemble members. Using a wide range of classification data sets, methodologies, and fusion techniques, current diversity research is extended by expanding classifier domains before employing fusion methodologies. The expansion is made possible with a unique classification score algorithm developed for this purpose. Correlation and linear regression techniques reveal that the relationship between diversity metrics and accuracy is tenuous and optimal ensemble selection should be based on ensemble accuracy. The strengths and weaknesses of popular diversity metrics are examined in the context of the information they provide with respect to changing classification thresholds and accuracies.

Keywords

Accuracy, classifier fusion, classification threshold, classification, diversity, ensembles

Introduction

There is considerable effort in the pattern recognition field to combine the outputs of individual classifiers to create a multiple classifier system (MCS), also termed an "ensemble," which endeavors for robustness over any single classifier in the MCS. The underlying principle is that greater accuracy can be achieved by combining the outputs of classifiers strong in different areas of the decision space. Classifiers that are strong in different areas of the decision space are said to be diverse, and intuitively selecting diverse classifiers would lend itself to improved accuracy.

However, the concept of diversity has yet to be formalized but there is consensus among researchers that diverse classifiers make errors in different areas of the classification domain. Herein, we consider diversity to be the abstract concept that describes differences between outputs of multiple distinct classifiers, while a diversity metric will be considered to be a rigorously defined method for describing this abstract concept of diversity. Currently, there are many proposed diversity metrics, c.f. literature,[1–3] without any clear consensus as to which diversity metric is best.

Although studies have examined the relationship between accuracy and diversity, c.f. literature,[4–6] limitations of these studies include that only a small part of the possible classification domain was considered. By selecting different classification thresholds for each individual classifier in an MCS, it is possible to look at a much wider range of the classification domain. We introduce an alternate scoring technique that allows selection of individual classification thresholds to generate a classification *surface* instead of just a single

Department of Operational Sciences, Air Force Institute of Technology (AFIT), Wright Patterson AFB, Dayton, OH, USA

Corresponding author:
Trevor J Bihl, Sensors Directorate, Air Force Research Laboratory (AFRL), Wright Patterson AFB, Dayton, OH 45433, USA.
Email: trevor.bihl.2@us.af.mil

classification *curve*. Through employing the alternative scoring technique, we find that the relationship between diversity and accuracy in ensembles is ambiguous, despite there is a statistically significant relationship between accuracy and diversity when using academic data sets, classification algorithms, and ensemble techniques. The alternate scoring technique allows us to create and evaluate a large number of MCSs, from which we analyze with linear correlation, least squares regression, as well as accuracy-based and diversity-based ensemble selection algorithms to uncover a possible relationship between accuracy and diversity.

This paper is organized as follows: the next section presents a review of fusion methods and diversity metrics. The subsequent section discusses underlying theory along with the proposed scoring method. Application results, from using academic datasets, are then presented. The concluding remarks are discussed in the last section.

Background

Prior research

Numerous studies have attempted to show a relationship between the diversity measures and the performance of an MCS.[3–16] Some studies have had some success in showing this relationship; however, they used diversity measures inherently correlated with accuracy. However, there have been no successes with the more "pure" measures of diversity.

Aksela and Laaksonen[4] studied classifier selection using a number of diversity metrics and fusion techniques and state that diversity metrics that disregard classifier correctness are not optimal for selection purposes. However, diversity metrics that take classifier correctness into account are "cheating" by really making the measure about accuracy instead of diversity. In essence, it is desirable for the diversity of the errors to be high, but the agreement on the correct outputs should also be high.[4]

This idea of diversity being important but not at the cost of accuracy is echoed in other research as well. Brown and Kuncheva[6] decomposed their diversity into "good" and "bad" diversity measures where increasing good diversity reduces error and increasing bad diversity increases error. However, they only did so for one fusion method and loss function combination; a separate decomposition must be performed for every combination of loss function and fusion method.[6] Brown and Kuncheva[6] also did not provide a way to use the good/bad diversity decomposition for building classifier ensembles. Canuto et al.[7]performed a study on ensemble selection with both hybrid (different types of classifiers) and non-hybrid (all classifiers are the same type) ensembles. They determined that classifier selection does have an impact on an ensemble's accuracy and diversity but they did not show any link between accuracy and diversity. They also show that hybrid ensembles provide the most diversity, this is one reason we use hybrid ensembles in our research. Gacquer et al.[8] proposed a genetic algorithm for ensemble selection that performs well with a specified accuracy-diversity trade off of 80/20, indicating that diversity must be of at least some use for selecting ensembles that generalize well over a population. However, they mentioned that this may not be true for small data sets, and it may not be true for all large data sets, either. Hadjitodorov et al.[9] looked at cluster ensembles which is a unsupervised learning technique, but still offer valid insight. They claim that accuracy peaks somewhere around medium diversity, and very high or very low diversity ensembles are a poor choice.

Alternatively, Kuncheva[10] stated that while no relationship between diversity and accuracy has been conclusively proven, it is may still be a useful idea in creating ensemble selection heuristics. Kuncheva and Whitaker[3] noted that the diversity metrics tend to cluster with themselves indicating that there is some agreed upon idea of diversity, but stated that using diversity for enhancing the design of ensembles is still an open question. Ruta and Gabrys[11] showed a correlation between one measure of accuracy, majority voting error, and two diversity metrics, the pairwise double-fault measure and the non-pairwise fault majority measure. The non-pairwise fault majority measure of diversity was designed specifically for majority voting fusion, and thus is expected to show a relationship with majority voting error.[11]

Shipp and Kuncheva[12] considered a large number of diversity metrics and fusion methods but did not find a correlation between ensemble accuracy and diversity. Windeatt[2] proposed a diversity metric that is measured across classes and not classifiers; he showed it to be correlated with the base classifier's accuracy but it did not appear to be correlated with the accuracy of the MCS as a whole. While some of the studies claim a correlation between accuracy and a proposed diversity metric, all of the studies fall short of conclusively proving a link between diversity and accuracy. Part of the problem stems from the fact that there is no formal definition of diversity.

With the current state of research in this area examined, one area that has not been researched at all is the relationship between accuracy and diversity over the classifier threshold domain space. All previous studies focused on the correlation between the classification accuracy at a fixed classification threshold, i.e. for a two class problem with a decision threshold $\theta = 0.5$, the class with posterior probability greater than 0.5 is the winning class. In this paper, the relationship

between diversity and accuracy is explored as the classification thresholds are varied over their full range, not just $\theta = 0.5$.

Classifier fusion

While it is possible to classify observations with a single classifier, greater accuracy may achieved by creating multiple classifiers and combining the results.[17] Combining multiple classifiers creates an MCS.

One of the most common structures is parallel combination, conceptualized in Figure 1, which we refer to as the standard method to contrast with our alternate method described later. The standard method is certainly not the only possible structure, and many other possibilities exist for combining classifiers. Fundamentally, all combination rules within the parallel structure fall into three different levels; an abstract level which only requires class labels as outputs, a rank level which requires a ranked list of class outputs, and finally a measurement level which requires class probabilities.[11]

Majority voting

Majority voting is the simplest abstract level fusion method. It involves selecting the most commonly assigned class as the final assigned class. If there is a case where no class gets more than one vote, the final assignment is given to the individual classifier with the best accuracy.[1] There are other possible voting methods than just the simple majority described above, c.f.,[1] but these are not used in our research.

Measurement level fusion

Measurement level fusion requires more information than abstract level fusion and possibly performs better due to the additional information over abstract level fusion methods. Measurement level fusion schemes require fuzzy measures on the interval [0, 1] as the classifier outputs. These fuzzy measures are treated as class probabilities or one of the other measures of evidence: possibility, necessity, belief, or plausibility. There are a wide range of measurement level fusion schemes, only some of the most popular are discussed below. The following symbol conventions are used with measurement level fusion:

- $\mu_j(x)$ – the support given by the MCS to class j for an observation x.
- $d_{t,j}(x)$ – the support given by the individual classifier t to class j for an observation x.
- T – the number of classifiers in the MCS

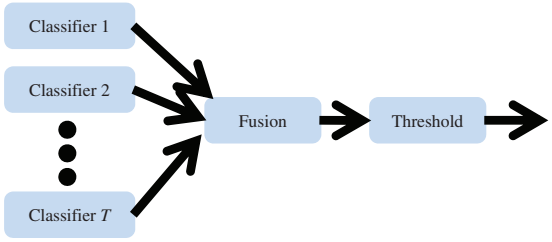

Figure 1. Conceptualization of the standard method.

Generalized mean

The generalized mean fusion method encompasses many commonly used fusion methods. The formula for a generalized mean fusion is[1]

$$\mu_j(x,\alpha) = \left(\frac{1}{T}\sum_{t=1}^{T} d_{t,j}(x)^{\alpha}\right)^{1/\alpha} \tag{1}$$

The choice of α determines the behavior of the rule. If $\alpha = 1$, we obtain the mean rule,[1] also called the basic ensemble model (BEM).[18] If $\alpha = -\infty$ then we obtain the minimum rule

$$\mu_j(x) = \min_{t=1\ldots T}\{d_{t,j}(x)\} \tag{2}$$

and similarly, $\alpha = \infty$ then we obtain the minimum rule,[1]

$$\mu_j(x) = \max_{t=1\ldots T}\{d_{t,j}(x)\} \tag{3}$$

Product rule

The product rule multiplies the support given by each classifier and if the posterior probabilities are correctly estimated then the product rule gives the best estimate of the overall class probabilities.[1] However, if one classifier gives very low support to a class, it effectively removes the chance of that class being selected

$$\mu_j(x) = \frac{1}{T}\prod_{t=1}^{T} d_{t,j}(x) \tag{4}$$

Generalized ensemble

The generalized ensemble model (GEM) is a generalized model of the mean rule, also called the BEM.[18] At its core, GEM is a weighted average of the support given by each classifier

$$\mu_j(x) = \sum_{t=1}^{T} \alpha_t d_{t,j} \tag{5}$$

The alphas are selected in a way that minimizes the mean squared error of the MCS. This is done by calculating the misfit function, $m_i(x)$, for each classifier

$$m_i(x) = f(x) - f_i(x) \tag{6}$$

where $f(x)$ is the truth and $f_i(x)$ is the output of classifier i. The correlation matrix between the misfit functions of all the classifiers i and j is then constructed, with individual entries

$$C_{ij} = E\left[m_i(x)m_j(x)\right] \tag{7}$$

The weights, α_i, are calculated using the entries in the correlation matrix

$$\alpha_i = \frac{\Sigma_j C_{ij}^{-1}}{\Sigma_k \Sigma_j C_{ij}^{-1}} \tag{8}$$

Perrone and Cooper[18] state that weights calculated using equation (8) creates the linear combination of classifier outputs that minimizes the MSE. GEM is proven to be more accurate than the best individual classifier and also more accurate than using BEM for fusing classifier outputs.

Diversity metrics

As discussed in the introduction, researchers do not agree on an exact definition of diversity or a definitive diversity metric. However, as mentioned by Polikar,[1] an effort must be made to make the component classifiers of an MCS as diverse as possible to ensure an efficient MCS. In the sections below, the most common measures of diversity are discussed, as well as how to use them to create a diverse set of classifiers. The challenge we face with current diversity measures is that the goal of linking diversity to accuracy is hampered by the fact that there is not a one to one mapping between diversity and accuracy. For each diversity metric discussed below, an example is provided to demonstrate how different sets of classifier outputs may have the same diversity but vastly different accuracies.

Diversity is easy to understand qualitatively, but difficult to rigorously quantify. There are many different measures that have been proposed to measure diversity. Some of the most popular metrics are discussed below. Most diversity metrics are designed for pairwise comparisons of classifiers. There are a few global diversity measures that can handle more than two classifiers such as Entropy and Kohavi–Wolpert Variance. A common approach is to compare multiple classifiers using pairwise diversity metrics by computing the diversity of every pairwise combination and averaging these results. In the pairwise diversity metrics, the convention used is the letters a, b, c, d represent fractions of instances as shown in Table 1.

Table 1. Reference for pairwise diversity metrics, from ChoiChaand Tappert.[19]

	Classifier j is correct	Classifier j is incorrect
Classifier i is correct	a	b
Classifier i is incorrect	c	d

Correlation

One of the most commonly used diversity metrics is the correlation between two classifiers, $\rho_{i,j}$.[3] Maximum diversity is obtained when $\rho_{i,j} = 0$. Correlation is calculated as

$$\rho_{i,j} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \quad 0 \leq \rho_{i,j} \leq 1 \tag{9}$$

Two identical classifiers that produce identical labels have $\rho = 1$ and fusing their outputs using BEM will give an MCS that has an accuracy equal to the accuracy of the individual classifiers. Another set of identical classifiers will also have $\rho = 1$, but if the accuracy of the individual classifiers in this new set does not equal the accuracy of the previously mentioned classifiers then the two MCSs will not have the same accuracy.

Yule's Q

Yule's Q statistic, $Q_{i,j}$ is another commonly used diversity metric, which takes on positive values if both classifiers tend to correctly classify the same instances, and negative values if both classifiers tend to incorrectly classify the same instances.[2] Maximum diversity is achieved at $Q_{i,j} = 0$. Yule's Q is calculated as:

$$Q_{i,j} = \frac{ad - bc}{ad + bc} \tag{10}$$

Two different MCSs can have the same Yule's Q statistic as long as the products ad and bc remain the same. For example, if one MCS has $a = 0.85\%$, $b = 0.05\%$, $c = 0.05$, and $d = 0.05$ and the other classifier has the same values except a and d have swapped values so $a = 0.05$ and $b = 0.05$ then both MCSs will have the same Yule's Q statistic but the first MCS will be very strong and the second MCS will be very weak.

Disagreement

Disagreement, $D_{i,j}$, is the probability that classifiers will disagree, and is calculated as[3]

$$D_{i,j} = b + c \tag{11}$$

Maximum diversity is achieved when $D_{i,j} = 1$. Two different MCSs can have the same disagreement but different accuracies as long as the sum $b+c$ remain the same. Similar to Yule's Q statistic, if the values a and d swap values, one MCS will be strong, while the other MCS will be weak even though they have the same disagreement.

Double fault

Double fault, $DF_{i,j}$ is the probability that both classifiers will misclassify an observation, and is equal to d[3]

$$DF_{i,j} = d \tag{12}$$

Maximum diversity is achieved when $DF_{i,j} = 0$. Two MCSs will have the same double fault value as long as they have equal values d. One MCS may have 99% "double correctness" and 1% double-faults, while another MCS may have 99% "single faults" and 1% double-faults. The former MCS is far more robust than the latter MCS despite them having the same double-fault values.

Entropy

Entropy, E, operates under the assumption that diversity is highest if half of the classifiers are correct and half of the classifiers are wrong. Diversity is highest when $E = 1$ and lowest when $E = 0$. Entropy is calculated as[1]

$$E = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T-\lceil T/2\rceil}\min(\zeta_i, (T-\zeta_i)) \tag{13}$$

where ζ_i is the number of classifiers that misclassified the observation x_i, therefore $(T-\zeta_i)$ is the number of classifiers that correctly classified observation x, and N is the number of observations in the data set. These definitions will also be used in the formula for Kohavi–Wolpert variance, discussed below. If one MCS always has three correct classifiers and two incorrect classifiers and the second MCS always has two correct classifiers but three incorrect classifiers then they will have the same entropy values but different accuracies.

Kohavi–Wolpert variance

Kohavi–Wolpert variance is similar to the disagreement measure but can be calculated with more than two classifiers. Diversity is maximized when Kohavi–Wolpert variance is high. Kohavi–Wolpert variance is calculated as[3]

$$KW = \frac{1}{NT^2}\sum_{i=1}^{N}\zeta_i(T-\zeta_i) \tag{14}$$

Kuncheva has proven that Kohavi–Wolpert variance of an MCS is related to the average of all pairwise disagreement.[3] Kohavi–Wolpert variance shares the same weaknesses as the entropy measure.

Methodology

Theory

Ruta and Gabrys[11] claim the difference between abstract level fusion techniques and measurement level fusion techniques is the information used by each technique, but there is one other difference that is important to this research. With an abstract level fusion method such as Majority Voting, class labels are given by each individual classifier then fused into a single label. Because class labels are given before the fusion takes place, each individual classifier can have its own decision threshold independent of the other classifiers. With a measurement level fusion method such as Mean Fusion, the measurements are fused and then a single label is made. Because there is only one label made (and it comes after fusion), there is only one decision threshold for the entire MCS. Although the measurement level fusion techniques make use of more information (fuzzy measures vs. binary labels), they lose degrees of freedom in that they cannot apply decision thresholds to individual classifiers. The following section proposes an alternate scoring technique that attempts to keep the increased information of the fuzzy measures required for measurement level fusion but allows each classifier output to be transformed independently.

Alternative scoring technique

The proposed alternate scoring technique, conceptualized in Figure 2, transforms class probabilities into scores restricted to the interval [0, 1] by selection of a classification threshold, θ.

The alternative scoring technique procedure takes classifier *t's* output probability of an observation belonging to class 1 $d^*_{t,1}$, and re-scores it to $d^*_{t,0}$ and $d^*_{t,1}$. The score not only captures the predicted class for an exemplar but also the relative distance of the original classifier score to

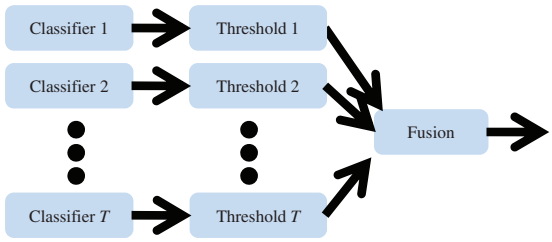

Figure 2. Conceptualization of proposed alternative scoring technique.

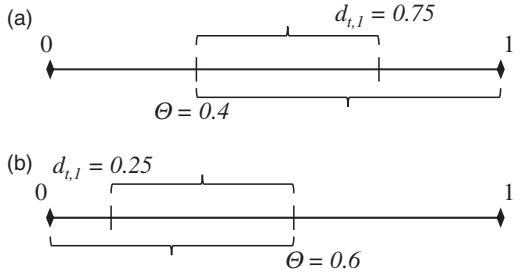

Figure 3. Graphical representation of proposed alternative scoring technique: (a) Using the alternate scoring technique, support for class 1 ($d^*_{t,1}$) is 0.35/0.6 = 0.58, support for class 0 ($d^*_{t,0}$) is 0 since $d_{t,1} > \Theta$. (b) A second example gives a support for class 0, ($d^*_{t,0}$) 0.45/0.6 = 0.75, support for class 1 ($d^*_{t,1}$) is 0 since $d_{t,1} < \Theta$.

the selected classification threshold

$$d^*_{t,0} = \max\left(0, \frac{\theta - d_{t,1}}{\theta}\right)$$
$$d^*_{t,1} = \max\left(0, \frac{d_{t,1} - \theta}{\theta}\right) \quad (15)$$

For an individual classifier, an assignment to class 0 would occur if $d^*_{t,0} > d^*_{t,1}$, and an assignment to class 1 would occur if $d^*_{t,0} \leq d^*_{t,1}$. A pictorial view of two examples is shown in Figure 3, once where $d_{t,1} > \theta$, and once where $d_{t,1} < \theta$. The alternate scoring technique will be applied to the classifier outputs prior to performing fusion, as opposed to the standard method which applies thresholds after performing fusion. The benefit of this is that it allows fusion methods employing the alternate scoring technique to look at many additional threshold combinations and explore a wider range of possible diversity and accuracy combinations. By allowing for individual classification thresholds, we can explore a greater range of diversity. To compare the procedural flow of the two methods, one can contrast Figures 1 and 2. Because the scores all fall on the same interval, we can perform the same fusion techniques on them as we could on class probabilities. We expect the performance of creating

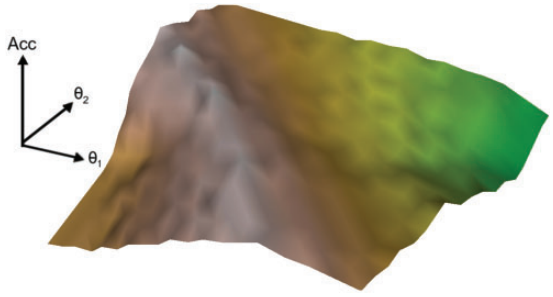

Figure 4. Sample accuracy surface over a range of thresholds.

ensembles using this alternate scoring technique to perform similarly to ensembles created using class probabilities. Mean fusion of ensemble alternate scores produces classification accuracy equal to mean fusion when all $\theta = 0.5$. A graphical comparison of the benefits of the alternate scoring technique is shown in Figure 3. Figure 4 shows an ensemble of two classifiers, with each point on this surface representing a single threshold combination, the height of the surface represents the accuracy. The alternate scoring technique can explore this entire domain, allowing a more in-depth look at the relationship between accuracy and diversity.

Experiment

The primary goal of this research is to discover if a relationship between ensemble accuracy and diversity exists.

Example academic datasets

In order to avoid data-driven results and to examine the relationship between accuracy and diversity across a wide spectrum of problem characteristics, 14 data sets were obtained from the UCI Machine Learning Repository.[20] The selected data sets: Balance Scale,[21] Breast Cancer Wisconsin,[22] BUPA Liver Disorders[23] Credit Approval,[24] Glass,[25] Haberman's Survival,[26] Fisher's Iris,[27] Mammographic Masses,[28] Parkinson's,[29] Pima Indians Diabetes,[30] Spambase,[31] SPECTF,[32] Transfusion,[33] and Wisconsin Diagnostic Breast Cancer.[34] These datasets have between 3 and 58 features, tens to thousands of observations, and benchmark accuracy values between 65% and 95%. All data sets have two classes or have been coerced into two class data sets by grouping similar classes until there are two distinct classes.

Classification algorithms

Six classifiers were employed to examine diversity and accuracy in ensembles:

- Quadratic discriminant analysis (QDA)

- k-Nearest Neighbors (kNN)
- Feed Forward Neural Network (FFNN)
- Radial Basis Function (RBF)
- Probabilistic Neural Network (PNN)
- Support Vector Machines (SVM).

Classifier settings and background were as follows:

1. QDA was considered, consistent with Wu et al.,[35] if a dataset was rank deficient then LDA, consistent with, Wu et al.[35] and Bihl et al.[36] was used.
2. kNN was employed consistent with Fukunaga and Narendra,[37] using the e1071 package[38] and default options.
3. FFNN was implemented per MeyerDimitriadou et al.[39] with one hidden layer with three nodes (used throughout), with a "softmax" (log-linear model).
4. RBF was implemented per Chen et al.[40] and DemuthBeale and Hagan,[41] with mean squared error goal of 0.0, spread = 1.0, max neurons equal to the number of input vectors, and 25 neurons added between displays.
5. PNNs were considered, per literature,[42–44] with a radial basis function spread of 0.1
6. SVMs used the e1071 package was used, c.f.,[38] with a linear kernel and default e1071 options.

For algorithms with tunable architecture settings, e.g. kNN, FFNN, RBF, PNNs, and SVMs, performance gains would logically be possible by selecting settings for each dataset. However, the authors have aimed for repeatability, consistent with the study in Liu and Zaidi,[45] in this study by using global settings which are likely overall suboptimal.

Experiment description

An area not examined in prior research and provided for by our alternate scoring technique is the relationship between accuracy and diversity over the entire domain of individual classifier thresholds. Most prior research has only investigated ensemble performance at single classification thresholds (typically $\theta = 0.5$). In the few studies where the thresholds were varied, only ROC curves of single classifier accuracy and MCS accuracy are presented. The ensembles we construct vary the classification threshold independently for each classifier and employ our proposed alternate scoring technique. Each unique combination of component classifiers and threshold settings produces a unique MCS whose accuracy and diversity may be examined. Another way to think of this experiment is as a full factorial design with the factors and levels as given in Table 2. This creates an experiment with 69,138,720 points. To evaluate the created ensembles, a function

Table 2. Experiment factor/level description.

Factor	# Levels	Notes
Data set	14	Previously mentioned data sets from UCI Machine Learning Repository
Fusion method	6	Maj. Vote, BEM, GEM, Product, Min, Max
Diversity metric	6	Correlation, Yule's Q, Disagreement, Double Fault, Entropy, Kohavi-Wolpert Variance
Classifiers	20	Out of six different classifiers (QDA, kNN, FFNN, RBF, PNN, and SVM), select 3 to make an ensemble
Thresholds	6859	19 thresholds (0.05 to 0.95 by 0.05) for each of the three classifiers in the ensemble

BEM: basic ensemble model; GEM: generalized ensemble model.

was created that takes as input the test and validation class probabilities from three classifiers, three individual classification thresholds as well as the truth. Using this information, the function performs the alternate scoring technique, calculates the diversity metrics, performs the fusion techniques, and returns the performance metrics of the fused ensembles. This function can be thought of as a wrapper that takes an ensemble and returns the desired performance metric, accuracy, and the desired diversity metrics; Correlation, Yule's Q, Disagreement, Double Fault, Entropy, and Kohavi–Wolpert Variance.

$$f\begin{pmatrix} test, validate, \\ class:1,\ \ class:2, class:3, \\ \theta_1, \theta_2, \theta_3, truth \\ = acc, \rho, Q, D, DF, E, KW \end{pmatrix} \tag{16}$$

In our experiments, every possible ensemble of three classifiers was evaluated at every threshold from 0.05 to 0.95 with threshold step sizes of 0.05. The diversity metrics and ensemble performances were saved in a database and used in the analysis performed.

Looking for relationships

There are a number of different ways to look for a relationship between accuracy and diversity with the wealth of data produced by our experimental design. One preprocessing step taken for all procedures was to map the diversity metrics to the interval [0, 1] where 0 is minimum diversity and 1 is maximum diversity. This mapping facilitates comparisons between accuracy and

diversity and allows their relative affects to be compared directly. Some diversity metrics already meet this criteria, such as disagreement and entropy. The remaining of the diversity metrics are mapped in the following manner

$$\begin{aligned} \rho^* &= 1 - |\rho| \\ Q^* &= 1 - |Q| \\ DF^* &= 1 - |DF| \\ KW^* &= 3 \cdot KW \end{aligned} \tag{17}$$

Correlations

The first logical step to uncovering a relationship between diversity and accuracy is to determine if there is a linear correlation between the diversity metrics collected and the ensemble accuracies. The correlation between test set diversity and test set accuracies are examined for within set correlation, and the correlation between test set diversity and validation set accuracies are examined for between set correlation.

Regression

Another possible way to uncover a relationship between diversity and accuracy is through linear regression. If there is a relation between diversity and accuracy then the validation set accuracy may be able to be predicted by test set diversity (which would be very useful in ensemble building). It is probable that test set accuracy is the main predictor of validation set accuracy and that diversity may only explain some of the residual error. To determine if this is the case, four regressions are performed on each data set- one with diversity as the only regressor, one with accuracy as the only regressor, one with both diversity and accuracy as regressors, and one with diversity and accuracy as regressors including their interaction. In each regression, the accuracy from the validation set is used as the dependent variable and all of the independent variables come from the test set. This ensures that the regressions show the actual predictive power of the independent variables and does not show spurious correlation within the test set. The regression results are examined to determine the effect of test set diversity and accuracy on validation accuracy. To account for the effects of the diversity metric used, the data set, the ensemble combination, and the fusion technique used, dummy variables are encoded. These dummy variables are included as main effects to allow for a change in the regression intercept, and are also interacted with testing accuracy, Acc_{TST}, and diversity, Div, to allow for the coefficients for accuracy and diversity to change based on the ensemble's properties (diversity metric, data set, and fusion technique). A regression of validation set accuracy with test set accuracy and diversity as the regressors without dummy variables is shown below

$$\widehat{Acc}_{val} = \beta_0 + \beta_1 Acc_{TST} + \beta_2 Div \tag{18}$$

This regression does not take into account the diversity metric, data set, and fusion technique in use. The full regression with dummy variables is

$$\begin{aligned} \widehat{Acc}_{val} = {} & \beta_0 + \beta_1 Acc_{TST} + \beta_2 Div + \beta_3 D_1 + \beta_4 D_2 + \beta_5 D_3 \\ & + \beta_{13} Acc_{TST} D_1 + \beta_{14} Acc_{TST} D_2 + \beta_{15} Acc_{TST} D_3 \\ & + \beta_{23} Div D_1 + \beta_{24} Div D_2 + \beta_{25} Div D_3 \end{aligned} \tag{19}$$

where D_1 is the vector of dummy variables associated with which diversity metric is used, D_2 is the vector of dummy variables associated with the data set the ensemble comes from, and D_3 is the vector of dummy variables associated with the fusion technique used. D_1 could be [00000] which would indicate the first diversity metric being used (correlation), a vector of [10000] would indicate the second diversity metric (Yule's Q), [01000] would indicate the third diversity metric being used (double-fault), etc. The other dummy variable vectors for data set and fusion technique are arranged similarly. The β_j' s and $\beta_{ij}'s$ associated with dummy variables are a vector as well. This full regression with dummy variables not only allows for the change of intercept and coefficients depending on the dummy variables and their interactions, it also allows for testing the statistical significance of the dummy variables and the information they are associated with.

Ensemble selection

To examine the utility of diversity to determine classifier membership in an ensemble, three ensemble selection schemes are used on the test set and compared against the most accurate ensemble and threshold combination in each validation set. The first scheme selects the ensemble with the highest ensemble test accuracy. The second scheme selects the ensemble with the three classifiers with the highest individual test accuracy. The third scheme selects the ensemble with the highest test diversity. These schemes are performed with each fusion type and their validation set accuracy is compared to the best ensemble's validation accuracy as determined by the oracle. These comparisons will be placed in percentages for relative comparison across fusion techniques, diversity measures, and data sets. If diversity is a useful metric to select classifiers for

an ensemble then the selection schemes that use diversity should compare favorably against the selection schemes that use accuracy.

Results

In our analysis, we evaluate the performance of the alternate scoring technique and ensure that it did allow us to look at a greater range of diversity. Next, we show the linear correlation between accuracy and the different diversity measures and the relative effects of accuracy versus diversity using regression techniques. Finally, we demonstrate the utility of selecting MCS membership using diversity as the primary criteria performs against using accuracy as the primary criteria. The experimental results first establish that the alternate scoring technique does provide ensemble selection over a wider range of diversity. Next, the correlation between accuracy and the examined diversity measures is examined. Following that outcome, the results from the regressions and the utility of using test accuracy and diversity to predict ensemble performance with validation data are presented.

Alternative scoring technique

The alternate scoring technique in general did not provide higher MCS accuracy but did allow examination of a greater range of diversity. For three of the fusion techniques: BEM, GEM, and Product Rule (denoted PRO in the tables), the alternate scoring technique was able to achieve a higher level of accuracy. With the two remaining fusion techniques, MIN and MAX, the alternate fusion technique did not achieve a very high level of accuracy. This is attributed to the manner in which the alternate scoring technique forces one of the scores to become zero which can greatly affect the behavior of these statistics. Table 3 shows a comparison of the alternate scoring technique's maximum and average performance for each fusion technique applied, averaged across all data sets. It is apparent the alternate scoring technique has the potential to perform as well as the standard method but loses some accuracy in the "tails" as the accuracy of the alternate scoring technique averaged across the range of classification thresholds is lower than the standard method. While we are pleased that the alternate scoring technique showed a potential improvement in "tuning" some ensemble techniques for better performance, the actual performance of alternate scoring technique is not of interest for this study. The primary reason for applying this technique is to allow us to examine a greater range of classification threshold combinations and a greater range of diversity. This focus on achieving greater diversity is why we feel that the lower average performance of the alternate scoring technique is acceptable.

Table 3. Comparison of standard method to alternative scoring technique-achieved accuracy.

Fusion	Max-Std	Max-Alt	Avg-STD	Avg-Alt
BEM	0.871	0.874	0.811	0.762
GEM	0.867	0.872	0.802	0.755
PRO	0.864	0.867	0.750	0.738
MIN	0.864	0.637	0.750	0.574
MAX	0.869	0.579	0.808	0.474

BEM: basic ensemble model; GEM: generalized ensemble model; PRO: product.

Table 4. Comparison of standard method to alternative scoring technique-achieved diversity range.

Metric	Range-Std	Range-Alt
Correlation	0.924	0.967
Yule's Q	0.955	0.982
Double-fault	0.408	0.424
Disagreement	0.549	0.653
Entropy	0.823	0.979
KW Variance	0.549	0.653

Diversity increase

Using the alternate scoring technique allowed the exploration of ensembles over a wider range of diversity. The expectation was that this greater range of diversity achieved would provide greater insight into the relationship between the accuracy and diversity of an MCS. As shown in Table 4 the alternate scoring technique achieves a higher range of diversity for every diversity metric. The diversity ranges are averaged across all data sets in Table 4. The use of the alternate scoring technique increased the diversity for every data set and all diversity metrics.

Ensemble combinations

The results of the experiment are described in Table 2.

Correlations

Similar to Kuncheva,[10] we begin our exploration of the relationship between diversity and accuracy by examining the correlation coefficient between the two measures. For each diversity metric and fusion method, we calculated the Pearson's r coefficient between the test diversity and test accuracy to determine if there was any within set correlation. The Pearson's r coefficient between the test diversity and validation accuracy was also examined to determine if there was any between

Table 5. Correlations by diversity metric.

Metric	Range-Std	Range-Alt
Correlation	0.023	−0.035
Yule's Q	0.352	0.238
Double-Fault	−0.106	−0.124
Disagreement	−0.106	−0.124
Entropy	−0.106	−0.124
KW Variance	0.001	−0.042

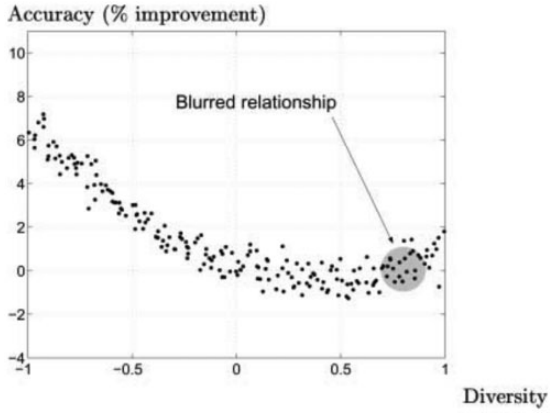

Figure 5. A typical accuracy-diversity scatterplot. Reprinted from Kuncheva.[10]

set correlation that could possibly be exploited for ensemble selection. The correlation aggregated by diversity metric is perhaps the most informative, and is presented in Table 5.

The correlation for all diversity metrics is small, and for most of the metrics, the sign is opposite what the conventional wisdom states. The conventional wisdom says that higher diversity should lead to higher accuracy and therefore have a positive correlation, but most of the correlation coefficients observed are negative. This result is supported, however, by Kuncheva[10] where she shows how for most of the diversity range there is a negative correlation with accuracy as shown in Figure 5, but once diversity exceeds a certain (fairly high) threshold, the relationship reverses to a positive correlation.

We believe that these results do not show anything new or novel; however, they serve to illustrate some common sense concepts about diversity. The more accurate a group of classifiers are, the less opportunity there is for diversity to exist. At the most extreme case if all the classifiers are 100% accurate then the ensemble will have zero diversity. Similarly, if all the classifiers are completely wrong then zero diversity will exist for all measures except for measures such as double-fault that only measure "half" the picture of diversity. We mentioned in the 'Diversity metrics' section, how each of the diversity measures we examined can produce multiple accuracy values for the same diversity value. Therefore, the results we observed are expected, there cannot be a one-to-one relationship between diversity and accuracy so no measure of correlation, linear or non-linear, will be able to show anything but a general trend. With regard to the double-fault measure, recall that double-fault measures the probability that both classifiers will misclassify an observation. We then changed to using the diversity score of $DF^* = 1 - |DF|$ so this now captures the probability that at least one classifier will correctly classify an observation. It should be clear then that this measure will have a positive correlation with accuracy.

Regression results

Regression analysis was performed to determine if a relationship exists between accuracy and diversity and can be used for ensemble selection. With this goal in mind, we use ensemble validation set accuracy as the response and metrics from the test set as the regressors. This process emulates a real-world application of picking an ensemble based on test set performance, with the validation set as new observations that are classified after an ensemble is selected. We performed three regressions; using test set accuracy as the only regressor, using test set diversity as the only regressor, and using both test set accuracy and diversity as regressors. Dummy variables were coded to allow for differences between data sets, fusion techniques, as well as the different diversity metrics. The primary focus was the coefficients related to accuracy and diversity, which gave insight on the relationship between accuracy and diversity. The results of the regressions are presented in Table 6, including the coefficients we were interested in as well as two measures of prediction performance (consistent with KutnerNachtsheim et al.[46]), the coefficient of determination (R^2) and root mean square error (RMSE). With over 69 million data points, practically any non-zero number will be statistically significant, c.f.,[47–50] thus statistical significance of the coefficients was not considered.

Readily apparent is that while diversity may be used as a selection criteria, diversity as the only regressor has the lowest R^2 and highest RMSE of the regressors examined. While an R^2 of 0.729 certainly indicates that diversity does offer some explanatory power for validation accuracy, it is outweighed by the much greater explanatory power of test set accuracy.

When both accuracy and diversity are included the linear model, the R^2 is increased only slightly, indicating that diversity does not provide much explanatory power beyond what accuracy already provides. Since all of the regressors are bounded on the same interval

Table 6. Regression coefficients + results.

Model	Acc	Corr	Yule's Q	DF	Disag	Entropy	KW	R^2	RMSE
Accuracy only	0.987	N/A	N/A	N/A	N/A	N/A	N/A	0.932	0.0404
Diversity only	N/A	0.043	0.045	0.223	−0.026	0.013	−0.026	0.729	0.0841
Accuracy + diversity	0.983	−0.005	−0.002	−0.008	−0.004	−0.001	−0.004	0.933	0.0402

RMSE: root mean square error.

Table 7. Percent achieved by fusion technique.

	Accuracy		Diversity					
Fusion technique	Ensemble	Individual	Corr	Yule's Q	DF	Disag	Entropy	KW
BEM	94	**95**	90	90	**93**	91	91	91
GEM	**93**	**93**	87	89	**91**	86	86	86
PRO	**95**	93	78	**80**	78	66	66	66
MIN	**95**	93	78	**80**	78	66	66	66
MAX	**95**	**95**	90	88	**91**	**91**	**91**	**91**
MVOTE	93	**95**	86	86	**93**	83	83	83

BEM: basic ensemble model; GEM: generalized ensemble model; PRO: product.

[0, 1], their coefficients can be directly compared to look at the effect of accuracy and each diversity metric. It is apparent that test set accuracy has a far greater impact on the validation set accuracy than any of the diversity metrics, indicating that even a large change in test diversity can only affect a small change in validation set accuracy. One observation to note is that of the diversity measures examined, the double-fault metric initially appears to be the best in terms of explanatory power. We believe this is again due to the fact that double-fault is more of a secondary measure of accuracy than it is a measure of diversity. As evidence of this, the coefficient for double-fault is relatively large compared to the other diversity measures when accuracy is not included in the regression, but when accuracy is included in the regression, the coefficient for the double-fault metric decreases to a level comparable to the other diversity measures.

Ensemble selection results

As a result of creating every possible ensemble combination, it was possible to determine which one of the possible ensembles was optimal for classifying each validation set. For each data set and fusion technique, there is an ensemble that delivers the maximum possible accuracy that can be obtained by choosing the very best combination of classifiers classification thresholds. We call these best possible ensembles "oracles" because that is the ensemble that an all-knowing oracle would select if it desired maximum performance. In our analysis, ensembles were selected based on results from the test set and the performance those ensembles achieved on the validation set was compared to the best ensemble selected by the oracle. Each selected ensemble's validation accuracy was compared to the oracle validation accuracy as a percentage

$$\%\ Achieved = \frac{Val.\ Acc.\ of given\ ensemble}{Oracle\ Val.\ Acc.} \tag{20}$$

The selection criteria used were ensemble test accuracy, individual classifier accuracy, and all six test diversity metrics. The percent performance that each selection criteria achieved, aggregated by fusion method, is shown in Table 7. In table 7, the best selection techniques based on accuracy and the best selection techniques based on the diversity measure are shown in bold.

As shown in Table 7, selecting ensembles based on accuracy achieves the highest performance for all fusion techniques, while selecting ensembles based on diversity gives lower performance. In fact, the lowest performing accuracy selection technique is never beaten (and is only tied once) by the highest performing diversity selection technique regardless of the fusion technique used. The double-fault diversity metric performed the best out of all the diversity metrics, but this is somewhat expected because of the inherent link between accuracy and the double-fault metric. This analysis shows that, even with an expanded range of diversity, test set accuracy should be the primary criteria for selecting ensembles. If there are two ensembles that tie in accuracy criteria, diversity

may be useful as a secondary criteria to break the tie, this will be investigated in further research.

Conclusions

This research presented an alternate scoring technique that allowed a wider range of diversity to be reached when creating MCSs. It demonstrated that there is not a one-to-one relationship between diversity and accuracy. Among the diversity measures examined, we single out the double fault metric as appearing to be the best measure, but this is likely due to its inherent link to accuracy and not due to it being a good measure of diversity. We have shown that validation accuracy is related to diversity but is greatly outweighed by the relationship between test set accuracy and validation accuracy. With our alternate scoring technique allowing us a wider range of ensembles to examine, we confirm that test set accuracy is still the best way to select ensembles.

Disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy of the United States Air Force, Department of Defense, or the U.S. Government.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Polikar R. Ensemble based systems in decision making. *IEEE Circuits Syst Mag* 2006; 6: 21–45.
2. Windeatt T. Diversity measures for multiple classifier system analysis and design. *Inform Fusion* 2005; 6: 21–36.
3. Kuncheva L and Whitaker C. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn* 2003; 51: 181–207.
4. Aksela M and Laaksonen J. Using diversity of errors for selecting members of a committee classifier. *Pattern Recog* 2006; 39: 608–623.
5. Brown G, Wyatt J, Harris R, et al. Diversity creation methods: a survey and categorisation. *Inform Fusion* 2005; 6: 5–20.
6. Brown G and Kuncheva L. "Good" and "bad" diversity in majority vote ensembles. In: *International workshop on multiple classifier systems*, Cairo, Egypt, 7–9 April 2010, pp. 124–133. Berlin, Heidelberg: Springer.
7. Canuto A, Abreu M, de Melo Oliveira L, et al. Investigating the influence of the choice of the ensemble members in accuracy and diversity of selection-based and fusion-based methods for ensembles. *Pattern Recog Lett* 2007; 28: 472–486.
8. Gacquer D, Delcroix V, Delmotte F, et al. On the effectiveness of diversity when training multiple classifier systems. In: *European conference on symbolic and quantitative approaches to reasoning and uncertainty*, Verona, Italy, 1–3 July 2009, pp. 493–504. Berlin, Heidelberg: Springer.
9. Hadjitodorov ST, Kuncheva LI and Todorova LP. Moderate diversity for better cluster ensembles. *Inform Fusion* 2006; 7: 264–275.
10. Kuncheva L. That elusive diversity in classifier ensembles. In: *Iberian conference on pattern recognition and image analysis,* Puerto de Andratx, Mallorca, Spain, 4–6 June 2003, pp. 1126–1138. Berlin, Heidelberg: Springer.
11. Ruta D and Gabrys B. Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems. In: *4th international symposium on soft computing*, Paper No. 1824-025, Paisley, UK, 2001, ISBN: 3-906454-27-4, pages 1–7
12. Shipp C and Kuncheva L. Relationships between combination methods and measures of diversity in combining classifiers. *Inform Fusion* 2002; 3: 135–148.
13. Didaci L, Fumera G and Roli F. Diversity in classifier ensembles: fertile concept or dead end? In: *International workshop on multiple classifier systems*, Nanjing, China, 15–17 May 2013, pp. 37–48. Berlin, Heidelberg: Springer.
14. Wang S and Yao X. Relationships between diversity of classification ensembles and single-class performance measures. *IEEE Trans Knowledge Data Eng* 2013; 25: 206–219.
15. Mohamad M, Saman MYM and Hitam MS. The use of output combiners in enhancing the performance of large data for ANNs. *IAENG Int J Comput Sci* Feb 13, 2014; 41(1): 38–47.
16. Mohamad M and Saman M. Comparison of diverse ensemble neural network for large data classification. *Int J Adv Soft Comput Appl* 2015; 7(3): 67–84.
17. Kuncheva LI. Diversity in multiple classifier systems. *Inform Fusion* 2005; 6: 3–4.
18. Perrone MP and Cooper LN. *When networks disagree: ensemble methods for hybrid neural networks*. In *How We Learn; How We Remember: Toward An Understanding Of Brain And Neural Systems: Selected Papers of Leon N Cooper*. Brown University, Providence, RI, pp. 342–358.
19. Choi S, Cha S and Tappert C. A survey of binary similarity and distance measures. *J Syst Cybernet Inform* 2010; 8: 43–48.
20. Lichman M. UCI machine learning repository, http://archive.ics.uci.edu/ml (2013, accessed 10 January 2018).
21. Siegler RS. Three aspects of cognitive development. *Cogn Psychol* 1976; 8: 481–520.
22. Mangasarian OL and Wolberg WH. Cancer diagnosis via linear programming. *SIAM News* 1990; 23: 1–18.
23. Forsyth RS. Liver disorders. In: *PC/BEAGLE user's guide*. Bupa Medical Research Ltd, 1990.
24. Quinlan JR. Simplifying decision trees. *Int J Man Mach Stud* 1987; 27: 221–234.

25. Evett I and Ernest J. Rule induction in forensic science, reading, Berkshire RG7 4PN: Central Research Establishment. Home Office Forensic Science Service, 1987.
26. Haberman SJ. Generalized residuals for log-linear models. In: *Proceedings of the 9th international biometrics conference*, 1976, pp. 104–122.
27. Fisher R. The use of multiple measurements in taxonomic problems. *Ann Eugen* 1936; 7: 179–188.
28. Elter M, Schulz-Wendtland R and Wittenberg T. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Med Phys* 2007; 34: 4164–4172.
29. Little M, McSharry P, Roberts S, et al. exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomed Eng Online* 2007; 6: 23.
30. Smith J, Everhart J, Dickson W, et al. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: *Proceedings of the symposium on computer applications and medical care*, Washington DC, 6–9 November 1988, pp. 261–265. Institute of Electrical and Electronics Engineers.
31. Hopkins M, Reeber E, Forman G, et al. *Spam base dataset*. Hewlett-Packard Labs, 1999.
32. Kurgan L, Cios K, Tadeusiewicz R, et al. Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artif Intell Med* 2001; 23: 149–169.
33. Yeh I-C, Yang K-J and Ting T-M. Knowledge discovery on RFM model using Bernoulli sequence. *Exp Syst Appl* 2009; 36: 5866–5871.
34. Street W, Wolberg W and Mangasarian O. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical Image Processing and Biomedical Visualization*, Vol. 1905. International Society for Optics and Photonics. San Jose, CA, 1993, pp. 861–871.
35. Wu W, Mallet Y, Walczak B, et al. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Anal Chim Acta* 1996; 329: 257–265.
36. Bihl TJ, Bauer K and Temple MA. Feature selection for RF fingerprinting with multiple discriminant analysis and using ZigBee device emissions. *IEEE Trans Inform Forensic Secur* 2016; 11: 1862–1874.
37. Fukunaga K and Narendra PM. A branch and bound algorithm for computing k-nearest neighbors. *IEEE Trans Comput* 1975; C-24: 750–753.
38. Meyer D, Dimitriadou E, Hornik K, et al. "ackage 'e1071', CRAN R Project., 2015.
39. Ripley B, Venables W and Ripley B. Package 'nnet'. R package version, 2016.
40. Chen S, Cowan C and Grant PM. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Trans Neural Netw* 1991; 2: 302–309.
41. Demuth H, Beale M and Hagan M. Radial basis networks (Chapter 7). In: *Neural network toolbox*. The MathWorks, Inc., 2008.
42. Wasserman P. *Advanced methods in neural computing*. New York: Van Nostrand Reinhold, 1993, pp. 35–55.
43. Situ J, Friend M, Bauer K, et al. contextual features and Bayesian belief networks for improved synthetic aperture radar combat identification. *Milit Oper Res J* 2016; 21: 89–106.
44. Specht D. Probabilistic neural networks. *Neural Netw* 1990; 3: 109–118.
45. Liu N and Zaidi NA. Artificial neural network: deep or broad? An empirical study. In: *Australasian joint conference on artificial intelligence*, Hobart, TAS, Australia, 5–8 December 2016, pp. 535–541. Berlin, Heidelberg: Springer.
46. Kutner M, Nachtsheim C, Neter J, et al. *Applied linear statistical models*. New York: McGraw-Hill Irwin, 2005.
47. Matthews JN and Altman DG. Interaction 2: compare effect sizes not P values. *BMJ* 1996; 313: 808–828.
48. Boos DD and Stefanski LA. P value precision and reproducibility. *Am Stat* 2011; 65: 213–221.
49. Anderson DR, Link WA, Johnson DH, et al. Suggestions for presenting the results of data analysis. *J Wildlife Manage* 2001; 65: 373–378.
50. Halsey LG, Curran-Everett D, Vowler SL, et al. The fickle P value generates irreproducible results. *Nat Meth* 2015; 12: 179–185.