# Optimising diversity in classifier ensembles of classification trees

Carina Ivaşcu, Richard M. Everson[0000−0002−3964−1150], and Jonathan E. Fieldsend[0000−0002−0683−2583]

University of Exeter, Exeter, UK
{ci233, R.M.Everson, J.E.Fieldsend}@exeter.ac.uk

**Abstract.** Ensembles of predictors have been generally found to have better performance than single predictors. Although diversity is widely thought to be an important factor in building successful ensembles, there have been contradictory results in the literature regarding the influence of diversity on the generalisation error. Fundamental to this may be the way diversity itself is defined. We present two new diversity measures, based on the idea of ambiguity, obtained from the bias-variance decomposition by using the cross-entropy error or the hinge-loss. If random sampling is used to select patterns on which ensemble members are trained, we find that generalisation error is negatively correlated with diversity at high sampling rates; conversely generalisation error is positively correlated with diversity when the sampling rate is low and the diversity high. We use evolutionary optimisers to select the subsets of patterns for predictor training by maximising these diversity measures on training data. Evaluation of their generalisation performance on a range of classification datasets from the literature shows that the ensembles obtained by maximising the cross-entropy diversity measure generalise well, enhancing the performance of small ensembles. Contrary to expectation, we find that there is no correlation between whether a pattern is selected and its proximity to the decision boundary.

**Keywords:** ensembles, classification, diversity, cross-entropy, hinge-loss

## 1 Introduction

A principal concern of supervised machine learning is to ensure a predictor demonstrates good *generalisation*. A predictor is considered to have the ability to generalise, if it has a good performance in predicting on unseen data drawn from the same process that it was trained on [1, 2]. Ensembles are collections of predictors, each of which is trained on a different subset of patterns or features. Some ensemble methods such as bagging [3] or boosting [4] have been seen to be very successful in pattern classification tasks [5], and ensembles have been proven in general to predict better than a single predictor [6, 7].

In this paper we consider classification of patterns $\mathbf{x}_n$, $n = 1, \ldots, N$ into two classes, the positive and the negative class. Each of the $M$ members of the

ensemble yields a score $y_{in} \equiv y_i(\mathbf{x}_n)$, $i = 1, \ldots, M$ indicating how likely it is that $\mathbf{x}_n$ belongs to the positive class, and the ensemble score $Y_n \equiv Y(\mathbf{x}_n)$, which may be converted to a decision by thresholding is, in general, the weighted average of the constituent predictor scores [8]:

$$Y_n \equiv Y(\mathbf{x}_n) = \sum_{i=1}^{M} c_i y_{in} \tag{1}$$

where $c_i$ are the non-negative weights assigned to the constituent ensemble members, $\sum_i^M c_i = 1$. Here we assume throughout that the ensemble members carry equal weight so that $c_i = 1/M$ for all $i$. When the constituent classifiers produce a hard decision and the weights are equal this amounts to the often used majority voting.

Various methods for assigning the classifier weights have been developed in [9–12]. Linear combinations have been mathematically investigated in [13, 14], together with nonlinear methods utilising rank-based information in [15], belief-based methods in [16–18] and voting schemes in [19, 20]. Here, however, we assume that the predictors are equally weighted and focus on the choice of patterns on which the ensemble members are trained.

Clearly, an accurate ensemble requires accurate members. However, Krogh and Vedelsby [21] have proven that an ensemble with good generalisation performance consists of members which disagree in their predictions [22]. As a result, diversity and accuracy are key factors in building successful ensembles.

Although the role of diversity has long been recognised, many ways of quantifying the diversity of an ensemble have been proposed. Kuncheva and Whitaker [23] empirically compared different diversity measures in order to assess the impact that diversity has on an ensemble's generalisation performance. However, their results could not support the influence of diversity on the overall performance of the ensembles. This aspect was partially explained in [24], which showed that different diversity measures have different degrees of correlation with generalisation error. It was also shown that there tends only to be high (negative) correlation between diversity and generalisation error when diversity is low and generalisation error is high; as diversity increases the correlation with generalisation error decreases [24]. We explore this aspect in more detail below.

In [21] Krogh and Vedelsby introduced a new diversity measure based on the ambiguity decomposition of regression ensembles and the bias-variance decomposition. The ambiguity term is obtained by subtracting the ensemble error from the average error of the predictors. Since the ambiguity is necessarily positive, this property shows the usefulness of the ensembles, since the ensemble error is lower than the average error of the classifiers. The ambiguity measures how much the predictions of the ensemble members differ from the ensemble prediction and as a result can be considered a type of diversity . Chen [24] defined another ambiguity measure in a similar fashion as to [21], but for classifiers and using the 0-1 loss. In his work, Chen demonstrated that out of all the diversity measures tested (Q-statistics, Kappa statistics, Correlation coefficient, Disagreement, Entropy, Kohavi-Wolpert variance, the measure of difficulty, generalised diversity,

coincident failure diversity), the ambiguity measure had the highest correlation with the generalisation error [24]. In this paper we use the term *ambiguity* to refer to a measure of ensemble diversity.

Here we further explore the connection between ensemble diversity and generalisation error. Following [21, 24], we define and characterise new ambiguity measures appropriate for the log loss and hinge loss. We investigate empirically the relationship between the ambiguity and the generalisation error. This leads to an evolutionary algorithm for the direct maximisation of the ensemble ambiguity, and thus generalisation error, by optimisation of the patterns that each ensemble member is trained on.

The principal contributions of our work are as follows:

1. the derivation of a cross-entropy-based ambiguity measure for ensemble diversity;
2. the derivation of a hinge-loss-based ambiguity measure for ensemble diversity;
3. the empirical assessment of the ambiguity/generalisation error trade-off on a number of widely used classification data sets, using decision trees ensembles;
4. the exploration of the effect of ensemble sampling rates on this trade-off;
5. the exploration of the direct *maximisation* of ensemble ambiguity via an evolutionary optimisation of the training patterns to maximise generalisation performance.

In the next section we present different diversity measures for ensembles using log and hinge losses. Section 4 presents an evolutionary algorithm for the optimisation of the cross-entropy diversity. Section 5 illustrates the performance of the evolutionary optimiser on a range of classification problems. Section 6 presents the conclusions and the future work.

## 2 Ambiguity measures

Extending the idea of quantifying diversity in regression ensembles [21], Chen [24] defined a new classifier ensemble diversity measure in terms of how diverse the outputs of the constituent classifiers are compared with the ensemble prediction. Following this line, we define new diversity measures as the difference between the average error of the individual classifiers forming the ensemble and the ensemble error; that is we define the ambiguity through the simple relation:

$$\text{Ensemble error} \quad = \quad \text{Average error} \quad - \quad \text{Ambiguity} \tag{2}$$

In line with [24], we call these measures of diversity *ambiguity measures*.

We first review the ambiguity for the 0-1 loss [24], before defining new ambiguities for the log loss and hinge loss.

### 2.1 Ambiguity measure for 0-1 loss

Here we assume that the targets, the true classes against which the classifiers are trained, are $t_n \in \{-1, +1\}$, $n = 1, \ldots, N$. Then the ensemble prediction for patterns $\mathbf{x}_n$ is

$$Y_n = \text{sign}\left(\sum_{i=1}^{M} c_i y_i(\mathbf{x}_n)\right) \tag{3}$$

and the error or loss for the ensemble classifying $\mathbf{x}_n$ is thus

$$L_{01}(Y_n \cdot t_n) = \begin{cases} 0 & \text{if } Y_n \cdot t_n \geq 0 \\ 1 & \text{if } Y_n \cdot t_n < 0. \end{cases} \tag{4}$$

We denote the outputs of the ensemble members when classifying patterns $\mathbf{x}_n$ by $\mathcal{Y}_n = \{y_{in} = y_i(\mathbf{x}_n)\}_{i=1}^{M}$. Then, using (2), the corresponding ambiguity in the ensemble when classifying a single $(\mathbf{x}_n, t_n)$ pair is thus [24]:

$$\text{amb}_{01}(\mathcal{Y}_n) = \frac{1}{2}\sum_{i=1}^{M}(\frac{1}{M}Y_n - c_i y_{in})t_n. \tag{5}$$

The ambiguity of the ensemble for a dataset of $N$ patterns is just the ambiguity for each pattern averaged over the $N$ patterns.

$$\text{amb}(\mathcal{Y}) = \frac{1}{N}\sum_{n=1}^{N}\text{amb}(\mathcal{Y}_n) \tag{6}$$

for the 0-1 loss and the other losses which we consider. It can be shown that (see Supplementary Material) the 0-1 ambiguity is zero if and only if all the ensemble members agree on the classification of a pattern, that is $\text{amb}_{01}(\mathcal{Y}_n) = 0 \Leftrightarrow\ y_{in} = y_{jn}\ \forall 1 \leq i, j \leq M$. We note, however, that $\text{amb}_{01}(\mathcal{Y}_n) < 0$ if $Y_n \neq t_n$ so that the ambiguity is negative if the ensemble classification is incorrect.

### 2.2 Ambiguity measure for log loss

The cross-entropy error or log loss measures the discrepancy between the output of the classifier and the true class when the classifier produces an output between 0 and 1 which may be interpreted as a posterior probability; for convenience we denote the classes as 0 and 1, $t_n \in \{0, 1\}$. We can express the loss for the $i$th classifier on the $n$th pattern as:

$$L_{log}(y_{in}, t_n) = -\left[t_n \log(y_{in}) + (1 - t_n)\log(1 - y_{in})\right] \tag{7}$$

where $y_{in}$ is the probability prediction of the $i^{th}$ classifier for the $n^{th}$ pattern belonging to the positive class. The error made by the ensemble for the $n$th pattern is therefore quantified as:

$$L_{log}(Y_n, t_n) = -[t_n \log(Y_n) + (1 - t_n)\log(1 - Y_n)]. \tag{8}$$

Again defining the ambiguity as the difference between the average loss of each member of the ensemble and the ensemble loss we obtain the cross-entropy ambiguity for a single pattern:

$$\text{amb}_{CE}(\mathcal{Y}_n) = \sum_{i=1}^{M} c_i L_{log}(y_{in}, t_n) - L_{log}(Y_n, t_n). \tag{9}$$

Using equations (7), (8) and (9), we obtain:

$$\text{amb}_{CE}(\mathcal{Y}_n) \triangleq t_n \log\left(\frac{\sum_{i=1}^{M} c_i y_{in}}{\prod_{i=1}^{M} y_{in}^{c_i}}\right) + (1-t_n)\log\left(\frac{\sum_{i=1}^{M} c_i (1-y_{in})}{\prod_{i=1}^{M}(1-y_{in})^{c_i}}\right). \tag{10}$$

Note that for any $t_n$ only one of the terms will not be zero, so $\text{amb}_{CE}(\mathcal{Y}_n)$ is the logarithm of the ratio between the arithmetic and geometric means of the proximity of the classifiers' outputs to the desired targets. The cross entropy ambiguity for many patterns is just the ambiguity averaged over patterns (6).

We note Woodhouse [25] shows that the ratio of the arithmetic mean to the geometric mean is equivalent to a cross-entropy quantifying the amount of information added in an image processing problem. In addition in [26] the ratio of the arithmetic to geometric mean is used to measure homogeneity.

Using the inequality between arithmetic and geometric means, namely that the arithmetic mean is greater than or equal to the geometric mean, it can be seen that $\text{amb}_{CE}(\mathcal{Y}_n) \geq 0$ for any input pattern. It can also be shown that $\text{amb}_{CE}(\mathcal{Y}_n) = 0$ if and only if all the constituent classifiers agree, $y_{in} = y_{jn}$ $\forall 1 \leq i, j \leq M$.

### 2.3 Ambiguity measure for hinge loss

Following the same route, an ambiguity measure can be obtained appropriate for the hinge loss. The hinge-loss is defined as:

$$L_H(y_{in}, t_n) = \max(0, 1 - t_n y_{in}). \tag{11}$$

Here $y_{in}$ is the $i^{th}$ classifier score for the $n^{th}$ pattern and $t_n$ is the target, where it is convenient to label the targets as $\{\pm 1\}$. The ambiguity measure obtained for the hinge loss is obtained by straightforward substitution, resulting in the following:

$$\text{amb}_{HL}(\mathcal{Y}_n) = \sum_{i=1}^{M} c_i \max(0, 1 - t_n y_{in}) - \max\left(0, \sum_{i=1}^{M} c_i (1 - t_n y_{in})\right). \tag{12}$$

As for $\text{amb}_{CE}$, the hinge loss ambiguity is non-negative: $\text{amb}_{HL}(\mathcal{Y}) \geq 0 \, \forall \mathcal{Y}$. However, while it is easy to verify that if all the component classifiers have the same score ($y_{in} = y_{jn}$ for all $1 \leq i, j \leq M$) then $\text{amb}_{HL}(\mathcal{Y}_n) = 0$, the converse is not true. This occurs when

$$1 - t_n y_{in} \geq 0 \, \forall i \in \{1, \ldots, M\}. \tag{13}$$

Inequality (13) can be satisfied when one of the component classifiers predicts incorrectly the class ($\exists i \in \{1, \ldots, M\}\, t_n y_{in} < 0$), whereas the others classify correctly the class, but with a score in absolute value lower or equal to 1 ($\forall j \in \{1, \ldots, M\}, j \neq i$, $t_n y_{jn} > 0$ and $|y_{jn}| \leq 1$). Proofs for the formulae of the ambiguity measures and their properties are presented in the Supplementary Material.

## 3 Correlation between ambiguity and generalisation error

Previous studies have investigated the relationship between diversity (measured in a variety of ways) and the error/loss [23, 24]. A negative correlation between generalisation error and ambiguity has been reported [24]. However, it is clear that this cannot be true across the entire range of ambiguity because it would imply that choosing the ensemble with the maximum diversity would minimise the generalisation error, but a maximally diverse ensemble (with no predictive power) could be constructed from learners that make random predictions. We therefore empirically investigate the relationship between the ambiguity measured on a *training* data set and the error/loss on a test data set (approximating the generalisation error).

Bagging was used in order to control the diversity by sampling different independent samples to train the classifiers in the ensemble. We use 30 sampling rates in the range $[0.01, 1]$. For each sampling rate an ensemble of decision trees, forming a random forest [3] was trained on the sampled patterns. From the 2000 available observations, 1000 were drawn at random and used for training, while the remaining 1000 for evaluating the generalisation error; the roles of the training and testing sets were then swapped and the corresponding ambiguities and losses calculated. This process was repeated 50 times and the ambiguities and errors averaged over the resulting 100 instances.

We used the GMM5 dataset [27] which comprises two-dimensional features generated by a Gaussian mixture model with 5 components (an extension of the 4-component model of [28]) allowing a large quantity of data to be synthesised and the Bayes error rate to be calculated exactly.

Figure 1 shows the variation of the generalisation error with the diversity of the ensemble measured on the training dataset for each of the ambiguity measures discussed. The first column of panels in Figure 1 corresponds to a small ensemble of $M = 5$ trees the second column shows the variation for a large ensemble of $M = 100$ trees. Although there is considerable variation between the curves for the different ambiguity measures, they all display common characteristics. At high sampling rates the ambiguity and test error are negatively correlated, as also reported by [24]. In this regime, as the sampling rate increases member classifiers are trained on increasingly similar views of the data and therefore diversity decreases. Since the average error per classifier is approximately constant (because adding more data does not appreciably increase their accuracy), equation (2) shows that the ensemble error increases.

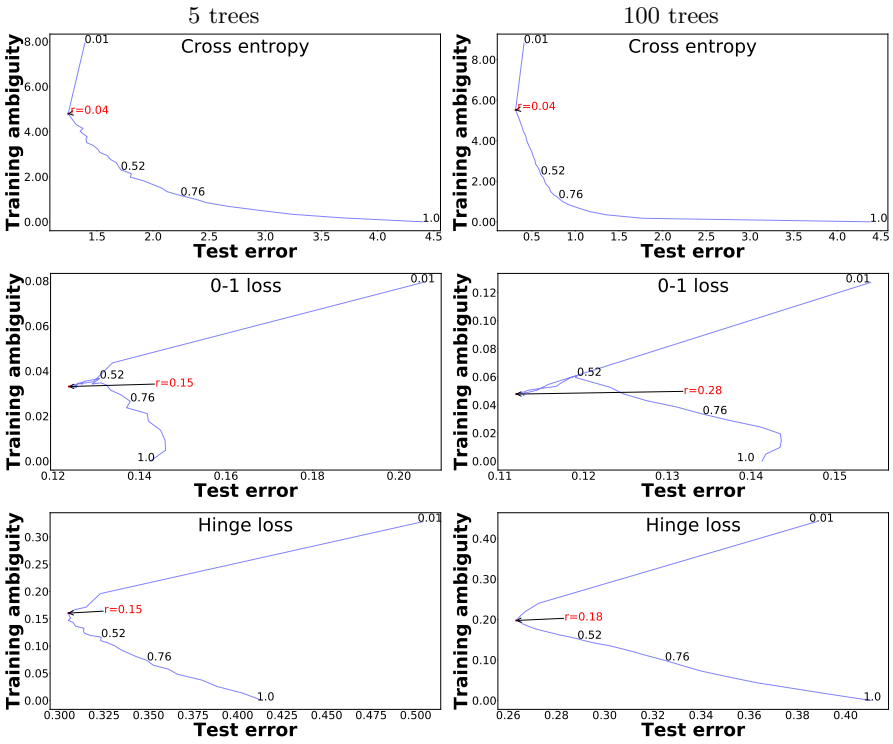


**Fig. 1.** Curves of the three types of ambiguities versus the corresponding losses that were derived from the ambiguity measures detailed in Section 2. The test error versus the training ambiguity was plotted for different sampling rates for ensembles formed of 5 trees (left column) and 100 tress (right column) for the gmm5test dataset. The first row shows the behaviour of the test cross entropy versus the training cross entropy ambiguity, in the second row the test 0-1 loss versus its corresponding training ambiguity is plotted, respectively the behaviour of the hinge loss is presented in the third row of panels. The optimal sampling rate ($r$) is indicated in red.

Decreasing the sampling rate means that the members of the ensemble are trained on different views of the data, leading to increasing diversity/ambiguity and therefore a smaller ensemble error c.f. (2). However, as the sampling rate is reduced to even lower levels, each component classifier is trained on a very small number of patterns and therefore starts to become inaccurate. In (2) the average error increases more rapidly than the diversity and the result is that the ensemble error begins to rise again. Unfortunately, determining the sampling rate that yields the best generalisation error is not straightforward or susceptible to *a priori* analysis. In section 4 we therefore describe an evolutionary algorithm to determine this rate.

The same pattern is apparent for both small ($M = 5$, Figure 1 left column) and large ($M = 100$, Figure 1 right column) ensembles, although the larger

ensemble achieves a lower generalisation error. This generalisation error is very close to the Bayes error (0.11 misclassification rate) for this data set. It might be expected that the optimum sampling rate would be at least $1/M$, so that each classifier in the ensemble is trained on $N/M$ examples and each example is used on average in the training of at least one classifier. However, as the panels in Figure 1, the optimum sampling rate is well below $1/M$, meaning that some of the data is not used at all by the ensemble. This indicates the significant role played by diversity: to achieve best generalisation performance it is better to ensure diversity by exposing classifiers to very different views of the data than to better train them by providing more data.

Although only shown here for the GMM5 dataset we emphasise that very similar relationships between ambiguity and generalisation error were observed on a number of additional datasets (Table 1). We also repeated the experiments using sampling with replacement, but bagging without replacement in general yielded lower generalisation errors.

We also investigated the variation of generalisation error with the number $M$ of classifiers forming the ensemble. This was achieved by generating ensembles with 2 to 100 members and training them, as before, with samples at a given rate. This was repeated 20 times for each ensemble size and sampling rate. The average (test) cross entropy error plotted against size of ensemble and sampling rate is shown in the panel of Figure 2 for the Sonar data set (Table 1, [29, 30]). This figure plainly shows the benefit of a large ensemble: the optimum generalisation error with a large ensemble is obtained over a wide range of sampling rates. The average training cross entropy ambiguity is plotted against size of ensemble and sampling rate in the right panel of Figure 2. These two figures together show the relationship between generalisation error and training ambiguity; high ambiguities yield lower test errors, provided the sampling rate is not too small. However, these two plots show the difficulty of predicting from the training ambiguity the optimal rate that will yield the lowest generalisation error.

## 4 An Evolutionary Algorithm to Optimise Ambiguity

As we have shown, provided that the sampling rate is not too low, the generalisation error is reduced for ensembles with high diversity. We therefore use an evolutionary algorithm to maximise the ambiguity of an ensemble of classifiers by selecting the patterns, that is the particular training examples, on which the constituent optimisers are trained . Pseudocode for the algorithm is presented in Algorithm 1.

We use ensembles of $M$ classifiers, each of which is trained on a fraction $\rho$ of the $N$ available training patterns. In common with standard bagging ensembles, each of the classifiers is trained on all the available features. The patterns on which each classifier is trained is represented by a string of $N$ 0s and 1s, where a 1 indicates that the corresponding pattern is used to train the classifier, so that there are exactly $[\rho N]$ 1s in each string and $[\cdot]$ indicates rounding to the nearest integer. The strings representing the training patterns are initialised

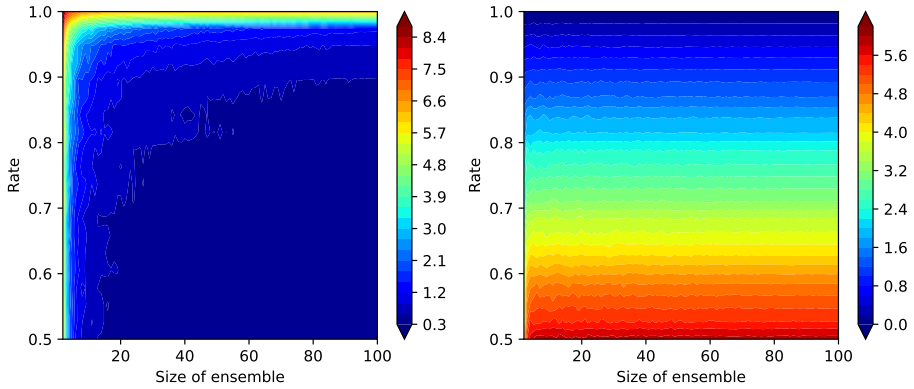


**Fig. 2.** The figure in the left of the panel represents the cross entropy generalisation error versus the size of the ensemble and the sampling rate. On the right hand side the training ambiguity derived from the cross entropy versus the size of the ensemble and the sampling rate is plotted. The plots were obtained for the Sonar data.

using stratified random sampling without replacement so that the class ratios are preserved.

A single ensemble is evolved through mutation. Between 1 and $M$ strings are mutated in one of two ways, chosen with equal probability (line 3 in Algorithm 1). Then a type of mutation is chosen with equal probability (line 5):

1. A proportion up to $\frac{N}{2}$ of 1s and 0s are flipped at random. This is performed in a stratified manner to preserve the class ratio and so as to maintain the sampling rate as $\rho$ (line 6).
2. The current string is discarded and replaced with a new string chosen in the same way as the initialisation, preserving the class ratio and the sampling rate (line 8).

Following mutation the $N_{pop}$ members with the largest ambiguity are retained to proceed into the next generation. In case of equality, the forest with the lower error will be preferred (line 10).

## 5 Experiments

We ran our algorithm on six standard classification datasets from the UCI Machine Learning Repository: Australian, Cancer, Liver, Heart, Sonar, Ionosphere [31] and an additional synthetic dataset GMM5 [28, 32]. Table 1 summarises the dataset characteristics.

Since the result shown in Fig. 2 show that for large ensembles, the generalisation error is small for sufficiently low sampling rates, we concentrate on small ensembles. We used ensembles of $M = 5$ trees, which were implemented by using the `DecisionTreeClassifier` function from the sklearn library [33] in Python and the ambiguity measure $\text{amb}_{CE}(\cdot)$ derived from the log loss (10).

**Algorithm 1** Evolutionary algorithm for evolving a diverse ensemble

**Input:** $X = \{\mathbf{x}_n\}_{n=1}^N$ ▷ training data
**Input:** $t = \{t_n\}_{n=1}^N$ ▷ targets
**Input:** $M$ ▷ number of trees
**Input:** $g$ ▷ number of generations
**Input:** $\rho$. ▷ sampling rate
**Output:** $\mathcal{T}$ ▷ evolved forest

$\mathcal{T} \leftarrow initialize(X, t, M)$ ▷ generate a random ensemble/forest
**for** $i = 1 \rightarrow g$ **do**
  $m \leftarrow random(1, M)$ ▷ choose m trees to be changed
  $indices \leftarrow indicesToChange(M, m)$ ▷ choose the indices of the $m$ trees
  **if** $U(0, 1) < 0.5$ **then**
    $\mathcal{T}' \leftarrow mutate(\mathcal{T}, indices, \rho)$ ▷ mutation type 1
  **else**
    $\mathcal{T}' \leftarrow genNewTrees(\mathcal{T}, indices, \rho)$ ▷ mutation type 2
  **end if**
  **if** $(\text{amb}_{CE}(\mathcal{T}') > \text{amb}_{CE}(\mathcal{T}))$ **or**
    $(\text{amb}_{CE}(\mathcal{T}') = \text{amb}_{CE}(\mathcal{T})$ **and** $L_{log}(\mathcal{T}', t) < L_{log}(\mathcal{T}, t))$ **then**
    $\mathcal{T} \leftarrow \mathcal{T}'$
  **end if**
**end for**
**return** $\mathcal{T}$

**Table 1.** Dataset characteristics

| Datasets | Patterns | Features |
|---|---|---|
| GMM5 | 1000 | 2 |
| Australian | 690 | 14 |
| Cancer | 569 | 10 |
| Liver | 345 | 6 |
| Heart | 270 | 75 |
| Sonar | 208 | 60 |
| Ionosphere | 351 | 34 |

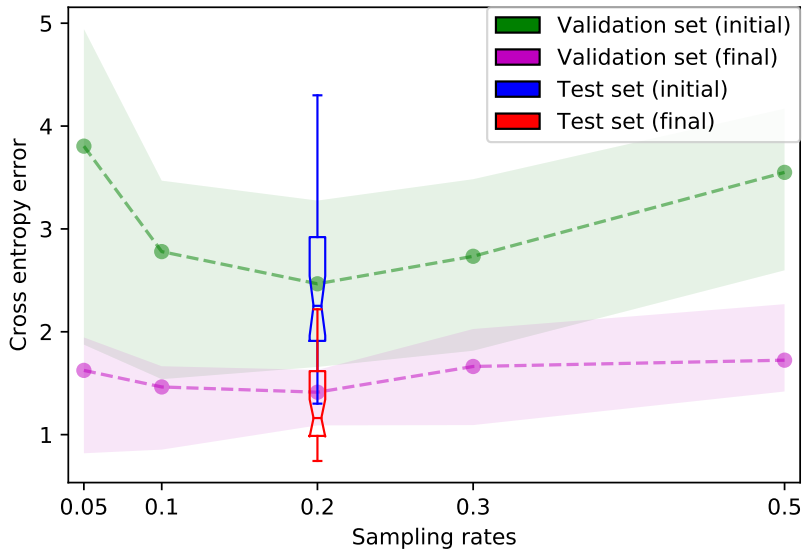


**Fig. 3.** Example results on the Liver dataset, using an evolutionary algorithm to optimise the cross-entropy ambiguity.

**Evolutionary algorithm** Data was partitioned into the following stratified parts as follows: one half for the test data, a quarter of the data for the training and the remaining quarter for the validation data. The evolutionary algorithm was run using the training data and the resulting ensemble evaluated on the validation data. The forest with the sampling rate that yields the lowest validation error was evaluated on the test data to assess the algorithm's performance.

Figure 3 shows example results obtained on the Liver dataset. The optimisation was repeated 30 times for each sampling rate and the figure shows the mean and interquartile range of the cross entropy generalisation error.

We compared the ensemble's validation error for the initial generation with the optimised ensemble's validation error, for the following sampling rates: 0.05, 0.1, 0.2, 0.3, 0.5. The green dashed line in Figure 3 corresponds to the mean of the 30 runs for the initial population, whereas the purple dashed line represents the mean for the final population. Shading indicates the interquartile range. The blue box plot corresponds to the test error for the initial populations, whereas the red box plots represents the test error for the corresponding final populations. These box plots were generated just for the sampling rate that yielded the lowest average validation error.

We also performed non-parametric statistical tests to assess the significance of the results. We used the Wilcoxon signed rank two-tailed test, $p = 0.05$. In Table 2 the mean test error of the initial ensemble for the sampling rate that yielded the smallest validation error is shown, along with the mean test error of the corresponding final evolved ensemble. The values in the parenthesis

**Table 2.** Results on datasets, mean over 30 runs given (lower and upper quartile in brackets). Bold mean value indicates significant difference (Wilcoxon signed rank two-tailed test, $p = 0.05$).

| Datasets | Initial cross entropy | Final cross entropy |
|---|---|---|
| GMM5 | 1.32 (0.82, 1.75) | **0.73** (0.6003, 0.852) |
| Australian | 1.35 (1.11, 1.52) | 1.26 (0.92, 1.39) |
| Cancer | 0.63 (0.35, 0.74) | **0.421** (0.32, 0.45) |
| Liver | 2.41 (1.91, 2.92) | **1.37** (0.98, 1.61) |
| Heart | 1.76 (1.195, 2.17) | **1.32** (0.94, 1.55 ) |
| Sonar | 2.19 (1.52, 2.953) | **1.21** (0.91, 1.52) |
| Ionosphere | 1.32 (0.97, 1.57) | **1.01** (0.83, 1.195) |

correspond to the 25th quartile and 75th quartiles. These results show that, in general, the EA performs significantly better than the random sampling from the initial population, and never worse. The ambiguity optimised ensembles have lower test errors on average than the initial ensemble across all test problems.

**What patterns are selected?** In our evolutionary algorithm we evolved the patterns that were selected in each tree. As such it would be interesting to see which patterns were actually chosen, and if they have any particular properties. In order to gain an understanding of which are the selected patterns, we analyse a two dimensional case.

A preliminary experiment was to plot the evolved patterns from the final generations of the evolutionary algorithm with their frequency of appearance. We performed this experiment just for the GMM5 dataset, because the distribution of these data are known and we have access to the posterior probabilities. We characterised the patterns according to their distance from the decision boundary. In order to determine how far a pattern is from the decision boundary, we calculated the maximum posterior probability of the pattern belonging to each of the two classes. The patterns belonging to the decision boundary have a minimum maximum posterior probability of 0.5. We averaged the number of appearances for the patterns from the final generation throughout the 30 runs. On the x-axis of Figure 4 the maximum of the posterior probability for both classes for each pattern is represented in 20 bins. On the y-axis, the proportion of occurrences is plotted. The green horizontal lines represent the medians of the number of occurrences for the patterns belonging to each of the 20 bins. This plot was obtained from the results of the evolutionary algorithm for the $\rho = 0.1$ sampling rate. Our results suggest that for this particular problem there is no preference for choosing some patterns during the optimisation, and that there is no correlation between whether a pattern is selected and its proximity to the decision boundary. This is contrary to what might be expected *a priori* — that is that points closer to the class boundary might be preferred as they give more information for bracketing the boundary.

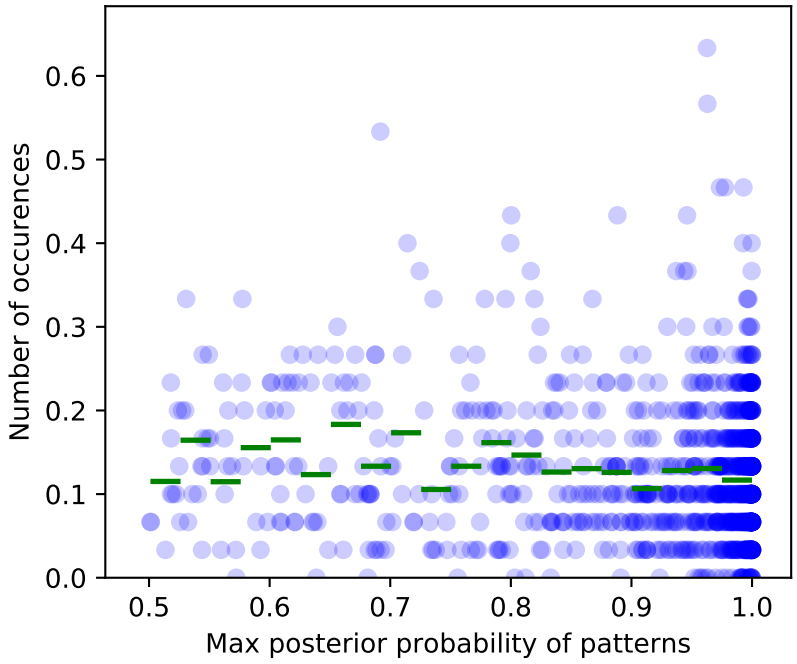


**Fig. 4.** Frequency of patterns selected by the evolutionary algorithm at the final generation for the gmm5test dataset, for the 0.1 sampling rate. On the x-axis is the maximum posterior probability of a pattern belonging to each of the two classes. The y-axis represents the average proportion each pattern was selected over the 30 runs of the evolutionary algorithm. The values from the x-axis have been divided into 20 bins. The green lines represent the medians of the number of occurrences of the patterns belonging to each bin.

## 6 Conclusion

In this paper we introduced two ambiguity measures using the bias-variance decomposition and the cross-entropy error or the hinge loss. Together with the ambiguity corresponding to the 0-1 loss, we established the properties of these new diversity measures. We evolved the training patterns of the classifiers in order to maximise the ambiguity obtained from the cross-entropy ($\text{amb}_{CE}$) and our results show that the evolved ensemble generally has a better generalisation error than the initial ensemble. Hence, our results support the influence that the diversity has on minimising generalisation error. Also the ambiguity measure obtained by using the cross-entropy error satisfies all the required properties of a diversity measure (being always positive and being zero if and only if the predictions of the classifiers are all the same). This property is not present in the ambiguity obtained by using the 0-1 loss (see [24]), which we find can be negative.

Our results show that if random sampling is used to select patterns on which ensemble members are trained, we find that generalisation error is negatively correlated with diversity at high sampling rates; conversely generalisation error is positively correlated with diversity when the sampling rate is low and the diversity high.

Also, we found that there is no correlation between whether a pattern is selected and its proximity to the decision boundary (at least for the problem we considered where we had direct access to the posterior probabilities and therefore could determine the 'true' decision boundary precisely).

Our experiments were based on random forests, therefore a possible extension of our work would be to use other types of ensembles and classifiers. In addition, other methods of inducing diversity, such as selection of features and different models, could be investigated.

In our experiments the weights $c_i$ of the classifiers were equal, as a result our future work will aim to optimise the weights of the classifiers in order to maximise ambiguity, without compromising the average error. Also some patterns have different ambiguities, so future work will focus on how to effectively select the most ambiguous patterns.

## References

1. Bishop C. *Pattern Recognition and Machine Learning*. Springer, 2006.
2. Duda R. O., Hart P. E., and Stork D. G. *Pattern Classification*. Wiley, New York, 2 edition, 2001.
3. Breiman L. Bagging predictors. *Machine Learning*, 24:123–140, August 1996.
4. Freund Y. and Schapire R. E. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, September, 1999.
5. Tang E. K., Suganthan P. N., and Yao X. An analysis of diversity measures. *Machine Learning*, 65:247–271, 2006.
6. Tumer K. and Ghosh J. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8:385–404, 1996.
7. Brown G., Wyatt J., Harris R., and Yao X. Diversity creation methods: A survey and categorisation. *Information Fusion*, 6:5–20, Mar 2005.
8. Tumer K. and Ghosh J. Error correlation and error reduction in ensemble classifiers. *Connect. Sci.*, 8:385–404, 1996.
9. Benediktsson J., Sveinson J., Ersoy O., and Swain P. Parallel consensual neural networks with optimally weighted outputs. *Proceedings of the World Congress on Neural Networks*, pages III:129–137, 1994.
10. Hashem S. and Schmeiser B. Approximating a function and its derivatives using mse-optimal linear combination of trained feedforward neural networks. *Proceedings of the Joint Conference on Neural Networks*, I:617–620, 1993.
11. Jacobs R. A., Jordan M. I., Nowlan S. J., and Hinton G. E. Adaptive mixtures of local experts. *Neural Computation*, 3:78–88, 1991.
12. Lincoln W. and Skrzypek J. Synergy of clustering multiple back propagation networks. *Advances in Neural Information Processing Systems*, 2:650–657, 1990.
13. Tumer K. and Ghosh J. Theoretical foundations of linear and order statistics combiners for neural patterns classifiers. *Technical Report 95-02-08, The Computer and Vision Research Center, University of Texas, Austin. (Available from URL http://www.lans.ece.utexas.edu/ under select publications-tech reports)*, 1995(c).

14. Tumer K. and Ghosh J. Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition*, 29(2):341–348, 1996.
15. Ho T.K., Hull J.J., and Srihari S.N. Decision combinations in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–76, 1994.
16. Rogova G. Combining the results of several neural networks classifiers. *Neural Networks*, 7(5):777–781, 1994.
17. Yang J.-B. and Singh M.G. An evidential reasoning approach for multiple-attribute decision making with uncertainty. *IEEE Transactions on Systems, Man and Cybernetics*, 24(1):1–19, 1994.
18. Xu L., Kryzyzak A., and Suen C.Y. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, 1992.
19. Hansen L. K. and Salamon P. Neural networks ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1000, 1990.
20. Battiti R. and Colla A.M. Democracy in neural nets: Voting schemes for classification. *Neural Networks*, 7(4):691–709, 1994.
21. Krogh A. and Vedelsby J. Neural network ensembles, cross validation and active learning. *Neural Information Processing Systems*, 7:231–238, 1995.
22. Chandra A. and Yao X. Multi-objective ensemble construction, learning and evolution. In *PPSN Workshop Multi-objective Problem Solving from Nature (Part 9th Int. Conf. Parallel Problem Solving from Nature: PPSN-IX)*, pages 9–13, 2006.
23. Kuncheva L. and Whitaker C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 2003.
24. Chen H. *Diversity and Regularization in Neural Network Ensembles.* PhD thesis, University of Birmingham, 2008.
25. Woodhouse I. H. The ratio of the arithmetic to the geometric mean: a cross-entropy interpretation. *IEEE Transactions on Geoscience and Remote Sensing*, 39(1):188–189, Jan 2001.
26. Chen C.C., Sen P.K., and Wu K.Y. Robust permutation tests for homogeneity of fingerprint patterns of dioxin congener profiles. *Environmetrics*, 23(285-294), 2012.
27. Fieldsend J.E., Bailey T.C., Everson R.M., Krzanowski W.J., Partridge D., and Schetinin V. Bayesian inductively learned modules for safety critical systems. *Computing Science and Statistics*, 35:110–125, 2003.
28. Ripley B. D. Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society Series B*, 56(3):409–456, 1994.
29. Gorman R. P. and Sejnowski T. J. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1:75–89, 1988.
30. UCI Machine Learning Repository. Connectionist bench (sonar, mines vs. rocks) data set. `https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks)`.
31. Dua D.and Graff C. UCI machine learning repository, 2017.
32. Krzanowski W. J., Fieldsend J. E., Bailey T. C., Everson R. M., Partridge D., and Schetinin V. Confidence in Classification: A Bayesian Approach. *Journal of Classification*, 23(2):199–220, 2006.
33. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., and Duchesnay E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.