Using Diversity in Preparing Ensembles of Classifiers Based on Different Feature Subsets to Minimize Generalization Error

Gabriele Zenobi, Pádraig Cunningham

Department of Computer Science
Trinity College Dublin
Gabriele.Zenobi@cs.tcd.ie
Padraig.Cunningham@cs.tcd.ie

Abstract. It is well known that ensembles of predictors produce better accuracy than a single predictor provided there is diversity in the ensemble. This diversity manifests itself as disagreement or ambiguity among the ensemble members. In this paper we focus on ensembles of classifiers based on different feature subsets and we present a process for producing such ensembles that emphasizes diversity (ambiguity) in the ensemble members. This emphasis on diversity produces ensembles with low generalization errors from ensemble members with comparatively high generalization error. We compare this with ensembles produced focusing only on the error of the ensemble members (without regard to overall diversity) and find that the ensembles based on ambiguity have lower generalization error. Further, we find that the ensemble members produced focusing on ambiguity have less features on average that those based on error only. We suggest that this indicates that these ensemble members are *local* learners.

1. Introduction

Ensembles of classifiers have recently emerged as a robust technique to improve the performance of a single classifier. Several ways to define an ensemble have been explored, from training each classifier in a subpart of the training set, to giving each classifier a subset of the features available.

When selecting an ensemble of classifiers a very simple approach consists of two separate steps: first a group of independently "good" classifiers is selected, then they are aggregated to form an ensemble. Such an approach has the advantage of simplicity, both conceptually and computationally, but the main disadvantage is that the classifiers are selected for the results they obtain singly and not for their contribution in the context of the ensemble. Following the work of Krogh and Vedelsby (1995), which demonstrated the crucial role played by the disagreement (ambiguity) in the final prediction of an ensemble, other less straightforward approaches have been proposed to build an ensemble of good predictors that have a high degree of disagreement. Among them the most relevant results were obtained by Liu (1999), who introduced a negative correlation penalty term to train ensembles of

neural networks, and that by Optiz and Shavlik (1996), who used the notion of ambiguity to find a diverse ensemble of neural networks using a genetic algorithm.

In this paper we focus on ensembles of classifiers based on different feature subsets and describe an algorithm that selects the different feature subsets (and thus the ensemble members) not just to minimize individual error but also to maximize ambiguity. This is compared with the default alternative of selecting the ensemble members based on error only without consideration for their contribution within the ensemble. In both scenarios the process of selecting the feature subsets is a "wrapper-like" search process (Kohavi & John, 1998) where Hill Climbing search is used to find a feature subset that minimizes error. In the default alternative (Cunningham & Carney, 2000) the search is guided by the error associated with the different feature subsets only. That research shows that the improvement due to the ensemble of nearest neighbour classifiers is correlated with the diversity in an ensemble. However, the diversity in the ensemble was determined after the ensemble was trained. Whereas, in the improvement presented here, the contribution of the ensemble member to the diversity of the ensemble is considered in the training process in order to ensure an ensemble of diverse members.

We present a study on ensembles of *k*-Nearest Neighbour (*k*-NN) classifiers that are trained on three different datasets with the two Hill Climbing approaches. The results show that the technique emphasizing ambiguity outperforms the strategy considering error only. Furthermore, we will see that forcing the classifiers to disagree leads to classifiers with a smaller number of features. This, as argued in (Cunningham & Zenobi, 2001) can be interpreted as an aggregation of several *local* specialists.

2. Ensembles and Diversity

The key idea in ensemble research is; if a classifier or predictor is unstable then an ensemble of such classifiers voting on the outcome will produce better results – better in terms of stability and accuracy. While the use of ensembles in Machine Learning (ML) research is fairly new, the idea that aggregating the opinions of a committee of experts will increase accuracy is not new. The Codorcet Jury Theorem states that:

> *If each voter has a probability p of being correct and the probability of a majority of voters being correct is M, then $p > 0.5$ implies $M > p$. In the limit, M approaches 1, for all $p > 0.5$, as the number of voters approaches infinity.*

This theorem was proposed by the Marquis of Condorcet in 1784 (Condorcet, 1784) – a more accessible reference is (Nitzan & Paroush, 1985). We now know that *M* will be greater that *p* only if there is diversity in the pool of voters. And we know that the probability of the ensemble being correct will only increase as the ensemble grows if the diversity in the ensemble continues to grow as well. Typically the diversity of the ensemble will plateau as will the accuracy of the ensemble at some size between 10 and 50 members.

In ML research it is well known that ensembling will improve the performance of unstable learners. Unstable learners are learners where small changes in the training data can produce quite different models and thus different predictions.

2.1 Different sources of diversity

Several ways to differentiate members of an ensemble of classifiers have been proposed in the literature. The most common source of diversity is by training members on different subsets of the training data. This can be done systematically by bootstrapping (sampling with replacement) different training sets from the training data. Such an approach has been applied with great success in eager learning systems such as Neural Networks (Hansen & Salamon, 1992) or Decision Trees (Breiman, 1996) This research shows that, for difficult classification and regression tasks, ensembling will improve the performance of unstable learning techniques such as Neural Networks and Decision Trees.

Ensembling will also improve the accuracy of lazy learners such as *k*-Nearest Neighbour (*k*-NN) classifiers, however *k*-NNs are relatively stable in the face of changes in training data so other sources of diversity must be employed. The popular solution is to use different feature subsets in the different classifiers (Ho,1998a;1998b), (Cunningham & Carney, 2000). This is the approach that we are concerned with here.

Other approaches such as different output targets and different learning hypothesis have also been considered.

Ensembles of different feature subsets

In this paper we focus on ensembles based on different feature subsets. The feature subset selection problem is very well studied in Machine Learning for a single classifier: it consists in finding the subset of features $F_S \subseteq F$ that maximizes performance. This is important when some of the features are irrelevant or redundant and consequently introduce some noise in the space of the instances. The main reasons why it is useful to perform feature subset selection are:

i) to build better predictors: better quality classifiers can be built by removing irrelevant and redundant features. This is particularly important for lazy learning systems;

ii) to achieve economy of representation and allow problems/phenomena to be represented as succinctly as possible;

iii) for knowledge discovery: to discover what features are and are not influential in weak theory domains.

A few studies have been done on the use of feature selection to create an ensemble of classifiers; among them those ones made by Cherkauer (1995), Ho (1998a, 1998b), Guerra-Salcedo and Whitney (1999a, 1999b) Tumer and Ghosh (1996) and Cunningham and Carney (2000) give the most promising results. However, if the use of ensembles improves the performance from one side, from another it reduces the other benefits of feature selection. It is clear that an ensemble of feature subsets affects the goal of economy of representation (ii) and also dramatically worsens the knowledge discovery (iii), mainly because we cannot say anymore that the outcome of a phenomenon depends on a particular subset of features. In the last section of this

paper we propose that the lack of interpretability associated with ensembles may be recoverable if the ensemble members prove to be *local* learners.

2.2 Different measures of diversity

There are a variety of ways to quantify ensemble diversity – usually associated with a particular error measure. In a regression problem (continuous output problem) it is normal to measure accuracy by the squared error so, as suggested by (Krogh & Vedelsby, 1995), a diversity measure can be variance, defined as:

$$a_i(x_k) = [V_i(x_k) - \overline{V}(x_k)]^2 \quad \textbf{(1)}$$

where a_i is the ambiguity of the i^{th} classifier on example x_k, randomly drawn from an unknown distribution, while V_i and $\overline{V}$ are, respectively the i^{th} classifier and the ensemble predictions. In this scenario the error from the ensemble is: $E = \overline{E} - \overline{A}$, where $\overline{E}$ is the average of the single classifier errors and $\overline{A}$ is the ambiguity of the ensemble. The equation also holds for classification, provided that the loss function used is the squared error function and that the ensemble prediction is still given as the weighted average of the single classifier predictions. Provided also, that we are happy to deal with real-valued class membership figures (see example below).

However, for classification the most commonly used error measure is a simple 0/1 loss function, so a measure of ambiguity in this case is:

$$a_i(x_k) = \begin{cases} 0 \text{ if } class V_i(x_k) = class\overline{V}(x_k) \\ 1 \text{ otherwise.} \end{cases} \quad \textbf{(2)}$$

where this time the classifier and ensemble outputs for the case labeled as x_k are classes instead of real numbers.

Another measure, associated with a conditional-entropy error measure, is based on the concept of entropy (Cunningham & Carney, 2000). This entropy measure of diversity can be defined as:

$$\tilde{A} = \frac{1}{M} \sum_{x=1}^{M} \sum_{k=1}^{K} - P_k^x \log(P_k^x) \quad \textbf{(3)}$$

This quantifies the overall entropy of an ensemble on a test set of M cases where there are K possible classes. While this is an useful measure of diversity it does not allow us to gauge the contribution of an *individual* to diversity so we will not use it here.

An Example

Our objective here is to identify an ambiguity measure that will help us determine the contribution of an individual ensemble member to diversity. The entropy-based measure (3) is not useful because it does not allow us to do that. The first two can

quantify the contribution of an individual member to ensemble diversity and the variance based measure has the advantage that it directly quantifies the improvement due to the ensemble. To see how these would be applied in practice a simple example is shown in Table 2 and Table 3. In order to use squared error and variance it is necessary that the outputs of the ensemble members are real valued. This is achievable in a variety of ways with nearest neighbour classifiers where a degree of class membership can be aggregated from the similarity to nearest neighbours.

Table 1. An example with 3 classifiers and 5 data points. The top half of the table shows ensembled predictions allowing continuous values and the bottom half shows 0/1 predictions

Value	**1**	**0**	**1**	**1**	**0**	E_i
Cl 1:pred	1	0.33	1	0.33	0.67	1
Cl 2:pred	0	0	0.33	0.67	0.33	1.67
Cl 3:pred	0.33	0.67	0.67	1	0	1
Ensemb.	**0.44**	**0.33**	**0.67**	**0.67**	**0.33**	**0.75**
Cl 1: 0/1	1	0	1	0	1	0.4
Cl 2: 0/1	0	0	0	1	0	0.4
Cl 3: 0/1	0	1	1	1	0	0.4
Ensemb.	**0**	**0**	**1**	**1**	**0**	**0.2**

Table 2. Error and ambiguity measures for the scenario shown in Table 2.

	E	$\overline{E}$	$\overline{A}$	$\overline{E}-\overline{A}$
Squared Err.	0.75	1.22	0.47	0.75
0/1 Loss	0.2	0.4	0.33	0.07

In the first scenario the outputs from the individual classifiers are real valued and we can see in Table 3 that the Ambiguity measure directly determines the improvement due to the ensemble as Krogh and Vedelsby predict. Using 0/1 loss and the ambiguity measure proposed above the ensemble still produces an improvement but it is not directly related to the ambiguity figure.

While the squared error and variance figures have this very elegant relationship these real valued class membership figures are not particularly meaningful so we will proceed using the 0/1 loss error measure and the ambiguity metric proposed in (2).

3. Using Ambiguity to select Ensembles of Classifiers

The aim of this study is to show how using ambiguity to select ensembles of classifiers will improve performance. One thing that appears to be clear is that to obtain good results an ensemble must include classifiers with a high degree of disagreement. It is this disagreement that gives the potential to correct the errors made by a single classifier. In the extreme case that all the classifiers are good but make mistakes over the same subset of data, the ensemble will not give a better performance than any single classifier.

To compare the default selection strategy, that doesn't take into account diversity, and ours, which makes use of diversity (ambiguity), we will use Hill-

Climbing search for a couple of reasons. First, we have a way to compare two ensembles performances that is not affected by any random event, once we state the same starting point (i.e. an initial set of feature masks). If we used for example a genetic algorithm it would be more difficult to make a direct comparison, due to its random nature. Second, a hill-climbing strategy is computationally less expensive then alternative stochastic search techniques. After all, we are interested in evaluating the heuristic to guide the search rather than the comprehensiveness of the search strategy.

3.1. The Default Search Strategy

In a classic hill climbing strategy *(HC)* that performs feature selection (Cunningham & Carney, 2000) a "good" classifier is selected by flipping each bit of the feature mask and accepting this flip if the classifier error E_i decreases. (A feature subset is a mask on the full feature set.) This process is repeated until no further improvements are possible – i.e. a local minimum in the feature set space is reached. The error is measured using leave-one-out testing. To produce an ensemble this process is repeated for each classifier and at the end all the classifiers are aggregated to form the ensemble. This approach is illustrated in Figure 1.

```
generate a random ensemble of feature subsets;

for every classifier i  in the ensemble  {
    calculate initial error E_i;
    do  {
          for every bit j of the mask  {
              flip j^th  bit of i^th  mask;
              calculate new E_i';
              if  E_i <= E_i'
                    flip back j^th  bit of i^th  mask;  //flip rejected
              else  E_i = E_i';  //flip accepted
          }
     }  while there are changes in the mask AND not  maximum number of iterations;
}

aggregate classifiers to obtain ensemble prediction;
```

Fig. 1. *HC:* the default selection strategy for generating ensembles using error only.

Clearly, from what we have said about the importance of diversity, this approach has the disadvantage that the improvement due to the ensemble may not be great because there is no means of promoting diversity in the ensemble.

3.2. *AmbHC:* a Hill-Climbing algorithm using Ambiguity

The dominant loss function used in classification is 0/1 loss and it is difficult if not impossible to derive a simple and linear equation that relates E to $\overline{E}$ and $\overline{A}$. However it is still clear that the *uplift* due to the ensemble depends on the diversity in the ensemble members (Cunningham & Carney, 2000). In the evaluation that follows we will use 0/1 loss and the associated ambiguity introduced in equation (2).

Assuming a homogeneous distribution of the instances (so that the average is simply obtained by dividing by *N*, the number of training samples) and equal weights in the ensemble, Ambiguity is defined as:

$$\overline{A} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{m}\sum_{i=1}^{m} a_i(x_k)$$

where $a_i(x_k)$ is given by equation (2). As the two summations are finite we can swap them, leading to the formula:

$$\overline{A} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{m}\sum_{i=1}^{m} a_i(x_k) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{k=1}^{N} a_i(x_k) = \frac{1}{m}\sum_{i=1}^{m} A_i$$

where the ambiguity A_i of the i^{th} classifier is defined as

$$A_i = \frac{1}{N}\sum_{k=1}^{N} a_i(x_k) \qquad \textbf{(4)}$$

On the basis of these definitions we may think of a new algorithm *(AmbHC)* that, taking the hill-climbing strategy as a starting point, tries to build an ensemble of classifiers with a high degree of disagreement. This approach considers every classifier in the context of the ensemble, and at each step accepts or rejects the flip depending on two parameters: the classifier error E_i and the classifier ambiguity A_i, as defined in the equation (4). If the improvement of one of the two parameters leads to a "substantial" deterioration of the other, then the flip is rejected. With "substantial" here we mean that a threshold value *(Thresh)* is given for the highest acceptable deterioration (if we consider acceptable a deterioration of 5% then *Thresh* takes the real value 0.05). The condition to accept or reject the flip of a bit is the following: if the improvement of one of the two parameters is less than the threshold value, then the highest acceptable deterioration of the other parameter is given by the improvement of the first one; if the improvement of one of the two parameters is instead greater than the threshold value, then the highest acceptable deterioration is the threshold itself. This technique allows us to avoid the selection of a set of good classifiers that make mistakes over the same subspace of the instances; it is illustrated in Figure 2.

In settling on this means of combining error and ambiguity in determining ensemble members we considered several alternatives; an evaluation of some of these is shown in Table 1. This table shows four columns of results for ensembles of size 13, 17 and 21. For the first column (HC (E)) the ensemble members were selected using error only. For the second, the selection was based on error minus ambiguity in the manner of Krogh and Vedelsby's (1995) work. The third is the same as the algorithm described in Figure 2 but without a threshold, i.e. flips are accepted if they improve ambiguity *or* error regardless of the effect on the other term. The fourth

column shows results for the algorithm shown in Figure 2. Clearly, the threshold approach work best. The E-A approach does not work so well because there is no basis for assuming that diversity has such a direct effect in classification. The technique without a threshold fails because sometimes improvements in ambiguity (or error) come at too high a cost in error (or ambiguity). Introducing the threshold overcomes this problem.

```
generate a random ensemble of feature subsets;

do {
 for every classifier i in the ensemble {
     calculate initial error E_i and contribution to ambiguity A_i ;
     for every bit j of the mask {
        flip j^th bit of i^th mask;
        calculate new E_i' and new A_i';
        if {  { [ E_i' < E_i ] AND
                 [    [( E_i' ≤ (1-Thresh ) × E_i ) AND (A_i' ≥ (1-Thresh) × A_i )] OR
                      [( E_i' > (1-Thresh ) × E_i ) AND (A_i' ≥ E_i'/E_i × A_i ) ]    ] } OR
              { [A_i' > A_i ] AND
                 [    [( A_i' ≥ (1+Thresh )× A_i ) AND (E_i' ≤ (1+Thresh )×E_i )] OR
                      [( A_i' < (1+Thresh )× A_i) AND (E_i' ≤ A_i'/ A_i ×E_i ) ]    ] }    }
        E_i = E_i' ; A_i = A_i';   //flip accepted
        else flip back j^th bit of i^th mask;  //flip rejected
     }
 }
} while there are changes in the masks AND not maximum number of iterations;

calculate final ensemble prediction;
```

Fig. 2. *AmbHC:* The algorithm for generating ensembles while emphasising diversity in ensemble members.

We have run the algorithms on three datasets; two available from the UCI repository (Pima Indians, Heart Disease) and the Warfarin data-set described in (Byrne et al., 2000). These were chosen on the basis of the following criteria:

- we have restricted our experimental comparison to 2-class datasets, turning a problem into a 2-class classification task if necessary, and have left the *n*-classe case for further research;
- we have considered datasets which do not have a skewed class distribution, as simple 0/1 error measures are questionable for datasets with very unbalanced class distributions.

In the next section we present a complete comparison of the results obtained by two of these algorithms, the basic hill-climbing error-only algorithm (*HC)* and *AmbHC*; below we give some further details about the *AmbHC* algorithm.

The *AmbHC* strategy is computationally comparable to the *HC* one. The only operation additional to *HC*, is the re-calculation at each step (i.e. every time we flip a bit of a classifier mask) of the classifier ambiguity A_i. This, in turn, needs a re-calculation of the whole ensemble prediction, but as the ensemble prediction is a linear combination of all the classifiers' predictions, at each step we only need to re-

calculate a single classifier prediction (any classifier prediction is not affected by a change in another classifier), just like it happens with a simple *HC*. Practically this means we need only to include a matrix that contains all the predictions, update it for any new value (in the row corresponding to that classifier) and recalculate two averages (ensemble prediction and ambiguity A_i). This adds a computational effort that is not problematic.

Table 1. Results of different alternatives for combining error and ambiguity in selecting ensemble members (the UCI Heart data was used).

	Ens Size	HC (E)	E-A	AmbHC (NoThresh)	AmbHC
Heart	13	17.7	18.8	17.8	17.2
	17	17.3	17.3	17.8	17.2
	21	17.7	18.1	19.7	16.9
Pima	13	25.0	25.7	26.0	24.5
	17	25.0	24.1	25.1	24.1
	21	24.6	24.7	25.8	23.8
Warfarin	13	7.8	7.9	8.1	7.8
	17	7.6	8.0	8.8	7.4
	21	8.0	7.4	8.0	7.3

4. Evaluation and Discussion

We present in this section a complete comparison of the *HC* and *AmbHC* selection strategies. We show that if the ensemble members are forced to be diverse then a better ensemble accuracy can be achieved with ensemble members that have poor overall accuracy, provided we include a sufficiently high number of classifiers in the ensemble. Also, these diverse ensemble members prove to have fewer features than ensemble members selected without consideration for diversity.

For each dataset we have run the two algorithms described in the previous section *(HC & AmbHC)*, varying their initial ensemble size. For each ensemble size we have also repeated the process with 4 different starting points (initial sets of feature masks), averaging the results obtained, as the hill-climbing strategy is quite sensitive to the initial condition. The scoring of any ensemble is determined using a 5-fold cross validation; in the 5-fold cross validation the data is divided into 5 parts and the ensemble is tested on each part in turn having been trained on the other 4 parts. The training involves the search processes described in Figures 1 & 2 and the fitness is determined using leave-one-out testing. The results are then averaged over the 5 validation sets. The threshold used for the *AmbHC* algorithm was set for all the datasets at 2.5%.

The evaluation on the three datasets shows that the ensembles trained with the *AmbHC* algorithm (higher diversity) have lower generalization errors than those trained with the simple *HC*, provided the size of the ensemble is sufficiently large (see Figures 3-5). Because of the nature of the ensemble training process *HC* ensembles have corresponding *AmbHC* ensembles allowing us to use a paired *t*-test to test the

hypothesis that the *AmbHC* ensembles have lower error. We have randomly selected 9 different ensembles in each of the three datasets and performed a paired *t*-test; the results gave a confidence of >80% for Warfarin, and >95% and >99% for Heart and Pima sespectively. These figures are very satisfactory - the weaker figure for Warfarin is probably accounted for by the small impact of the ensemble given the already low error of the individual classifiers. This is the first main result of our study: the algorithm that takes into account diversity while selecting ensemble outperforms the simply error-only strategy.

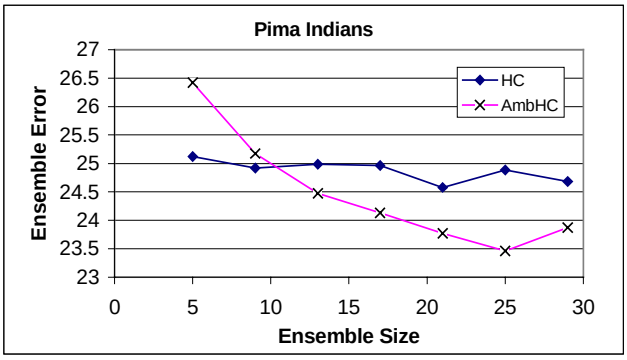

Fig. 3. Generalisaton error of different ensemble sizes on the Pima data..

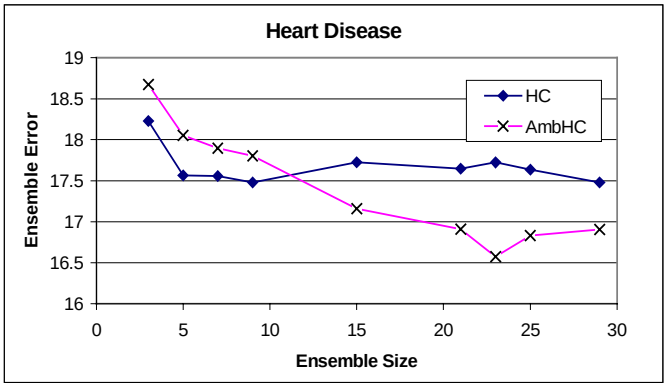

Fig. 4. Generalization error of different ensemble sizes on the Heart data.

In the following two tables we focus on some other aspects of the comparison between classifiers trained with the two different algorithms. In Table 4 we show, for each dataset and for both the algorithms, respectively the error obtained by the best ensemble, the average error and the average ambiguity of the single classifiers in the ensembles.

In Table 5 we show instead, for each dataset, respectively the total number of features and the average number of features of the masks trained with *HC* and with *AmbHC*.

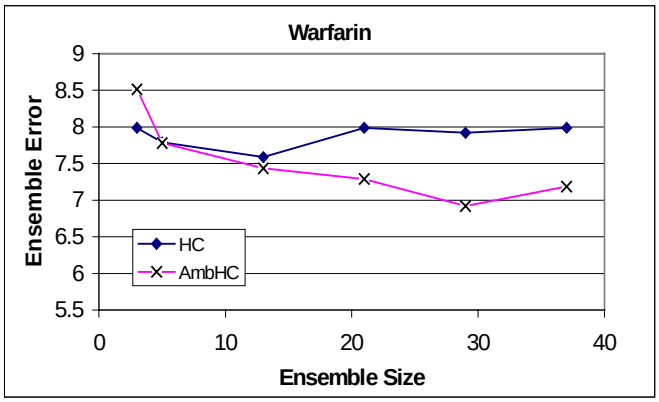

Fig. 5. Generalization error of different ensemble sizes on the warfarin data.

Table 4. A summary of all the evaluations showing the best ensemble generated for each data set and showing the corresponding average member error and ambiguity.

Data	Algorithm	Best Ensemble	Average Error	Avg. Ambiguity
Pima	*HC*	24.6	27.0	13.7
	AmbHC	23.4	31.5	22.3
Heart	*HC*	17.4	21.8	12.2
	AmbHC	16.6	24.8	18.8
Warfarin	*HC*	7.6	8.6	3.3
	AmbHC	6.9	14.1	10.6

Table 5. The ensembles built using *AmbHC* have significantly less features on average that those built using the default search algorithm.

Data	Total	Average: *HC*	Average: *AmbHC*
Pima	8	4.8	3.5
Heart	13	7.4	6.0
Warfarin	22	12.9	10.7

As we can see from Table 4, the classifiers in the ensemble selected with the *AmbHC* algorithm have a higher average error than those selected by the simple *HC* algorithm: the increase in ambiguity (diversity) comes at the cost of significantly higher errors in the ensemble members. It seems to us that the only way to account for the improvement in overall performance in the face of deterioration of the ensemble

members is that the members are local specialists. In fact, the use of ambiguity in the *AmbHC* algorithm means that the ensemble selection is made by choosing classifiers that disagree on a higher number of elements compared to those ones selected without ambiguity. For the first ones, the higher average ambiguity seems to compensate their higher error by 'distributing' the prediction of the different individuals over more diverse regions of the space of the instances; as a result we get a lower ensemble error.

This view is reinforced by another interesting result: the ensemble members produced using ambiguity have fewer features on average than the others (Table 5). It seems reasonable that fewer features are required to discriminate in these local regions. This observation also suggests a new perspective on the third issue connected to the feature selection problem (shown in the paragraph 2.1). It may be useful to reconsider this as a process of finding the best ensemble of local feature subsets rather than a global feature subset. However, the problem remains that if the ensemble is performing a problem space decomposition then it is doing so implicitly and the decomposition is not accessible. An interesting avenue for future research will be to use clustering to see if meaningful regions of the problem space can be identified where ensemble members specialize.

The local learners hypothesis helps us also in explaining the behaviour shown in the Figures 3 - 5, where, in each case, a minimum number of classifiers are needed for *AmbHC* to surpass *HC.* Since classifiers trained with the *AmbHC* algorithm have a higher average error than the ones trained with *HC* it is clear that each classifier will cover (i.e. predict correctly) a 'smaller' region of the problem space. So, to have the majority of the ensemble voting correctly we need a larger number of classifiers. Then, when the ensemble size is small (e.g. 5) even though we do not use diversity (in the *HC* algorithm) it is very probable that a set of classifiers randomly chosen has its own 'natural' diversity. As we increase the ensemble size, it becomes more probable that new members of *HC* ensembles will be similar to existing ones. While with the *AmbHC* algorithm diversity is still emphasized and variety is maintained. Thus, a diverse bunch of 'mediocre' classifiers outperforms a good bunch of classifiers with less diversity.

However, as mentioned in section 2 when discussing the Condorcet Jury Theorem, this addition of new diverse members does not continue to deliver benefit indefinately. At best, it will not continue to be possible to find more diverse members and the reduction in error will bottom out. What is happening in the three examples here is actually slightly worse than that in that an overfitting effect is evident. Since the selection of the ensemble members is a training process there is the potential for the ensemble as a whole to overfit the training data and that is clearly evident in the three graphs shown here. So with this *AmbHC* approach there is an optimal ensemble size which apperars to be between 25 and 30 for these data sets. It seems that the best way to address this overfitting would be to use a cross validation process to determine a best ensemble size.

5. Conclusions and Future Work

In this paper we have compared two approaches for selecting an ensemble of classifiers: a simple error-only strategy, where a group of independently 'good' classifiers is first selected and then aggregated, and a strategy which considers, during the training process, every classifier in the context of the ensemble and selects a group of classifiers with a high degree of diversity. We have focused our attention on ensemble of classifiers where diversity derives from different ensemble members using different feature sub-sets.

Since the objective of the evaluation has been to assess the feature selection strategies the comparison has been done using a simple hill-climbing search strategy. The strategies could be employed in a more comprehensive search algorithm such as a genetic algorithm or beam search.

Since there is a clear trade-off between diversity and error in the selection of the ensemble members the key question was; would diverse ensembles of (possibly) less accurate classifiers outperform ensembles of good classifiers with perhaps less diversity. The answer to this has proved to be 'yes' but it does depend on the careful management of the tradeoff between error and ambiguity that is implemented in the *AmbHC* algorithm as described in Figure 2.

This is interesting because it highlights something of a paradox associated with ensembles. It shows that it can be a good thing to have a committee of experts consistently voting 5 : 4 in favour of a prediction rather than 8 : 1. In fact, we are proposing selecting experts in a manner that will push down consensus in the committee. Intuitively, this is not what you want from a committee of physicians discussing your particular illness! You would like the committee of physicians to agree. A resolution of this paradox is as follows. If the committee members are very accurate there is little benefit in diversity; indeed there is little benefit in ensembles in classification tasks where accuracies of >93% are achievable with a single classifier. However, ensembles make sense where individual classifiers have significant errors (say > 15%). In such cases, instead of adding a new very accurate committee member that makes the same errors as existing members in the ensemble it is sensible to add a member that makes different errors, one that has a different set of competences. There is no benefit in adding members that will change votes of 8:1 to 9:1.

Perhaps the most interesting finding of this research is the fact that the ambiguity-focused learners have less features and the implication that these ensemble members are local learners. This may prove useful in understanding the contribution of ensembles in reducing error and may lead to an increase in the interpretability of ensembles. This will be the subject of our future research.

6. References

Breiman, L., (1996) Bagging predictors. *Machine Learning,* 24:123-140.

Byrne, S., Cunningham, P., Barry, A., Graham, I., Delaney T., Corrigan, O.I., (2000) Using Neural Nets for Decision Support in Prescription and Outcome Prediction

in Anticoagulation Drug Therapy, N. Lavrac, S. Miksch (eds.): *The Fifth Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2000)*.

Cherkauer, K.J. (1995) Stuffing Mind into Computer: Knowledge and Learning for Intelligent Systems. *Informatica* 19:4 (501-511) Nov. 1995

Condorcet, Marquis J. A. (1781) Sur les elections par scrutiny, *Histoire de l'Academie Royale des Sciences*, 31-34.

Cunningham, P., Carney, J., (2000) Diversity versus Quality in Classification Ensembles based on Feature Selection, *11th European Conference on Machine Learning (ECML 2000)*, Lecture Notes in Artificial Intelligence, R. López de Mántaras and E. Plaza, (eds) pp109-116, Springer Verlag.

Cunningham, P., & Zenobi, G., (2001) Case Representation Issues for Case-Based Reasoning from Ensemble Research, *submitted to ICCBR 2001.*

Guerra-Salcedo, C., Whitley, D., (1999a). Genetic Approach for Feature Selection for Ensemble Creation. *in GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference*, Banzhaf, W., Daida, J., Eiben, A. E., Garzon, M. H., Honavar, V., Jakiela, M., & Smith, R. E. (eds.). Orlando, Florida USA, pp236-243, San Francisco, CA: Morgan Kaufmann.

Guerra-Salcedo, C., Whitley, D., (1999b). Feature Selection Mechanisms for Ensemble Creation: A Genetic Search Perspective, in *Data Mining with Evolutionary Algorithms: Research Directions. Papers from the AAAI Workshop*. Alex A. Freitas (Ed.) Technical Report WS-99-06. AAAI Press, 1999.

Hansen, L.K., Salamon, P., (1990) Neural Network Ensembles, *IEEE Pattern Analysis and Machine Intelligence*, 1990. **12**, 10, 993-1001.

Ho, T.K., (1998a) The Random Subspace Method for Constructing Decision Forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 8, 832-844.

Ho, T.K., (1998b) Nearest Neighbours in Random Subspaces, *Proc. Of 2nd International Workshop on Statistical Techniques in Pattern Recognition*, A. Amin, D. Dori, P. Puil, H. Freeman, (eds.) pp640-648, Springer Verlag LNCS 1451.

Kohavi, R. & John, G.H., (1998) The Wrapper Approach, in *Feature Selection for Knowledge Discovery and Data Mining*, H. Liu & H. Motoda (eds.), Kluwer Academic Publishers, pp33-50.

Krogh, A., Vedelsby, J., (1995) Neural Network Ensembles, Cross Validation and Active Learning, in *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. S. Touretsky, T. K. Leen, eds., pp231-238, MIT Press, Cambridge MA.

Liu Y., Yao X. (1999) Ensemble learning via negative correlation, *Neural Networks* 12, 1999.

Nitzan, S.I., Paroush, J., (1985) *Collective Decision Making*. Cambridge: Cambridge University Press.

Opitz D., Shavlik J., (1996) Generating Accurate and diverse members of a Neural Network Ensemble, *Advances in Neural Information Processing Systems*, pp. 535-543, Denver, CO. MIT Press. 1996.

Tumer, K., and Ghosh, J., (1996) Error Correlation and Error Reduction in Ensemble Classifiers, *Connection Science, Special issue on combining artificial neural networks: ensemble approaches*, Vol. 8, No. 3 & 4, pp 385-404.