Diversity versus Quality in Classification Ensembles based on Feature Selection

Pádraig Cunningham, John Carney

Department of Computer Science

Trinity College Dublin
Padraig.Cunningham@tcd.ie

TCD-CS-2000-02

Abstract

Feature subset-selection has emerged as a useful technique for creating diversity in ensembles – particularly in classification ensembles. In this paper we argue that this diversity needs to be monitored in the creation of the ensemble. We propose an entropy measure of the outputs of the ensemble members as a useful measure of the ensemble diversity. Further, we show that using the associated conditional entropy as a loss function (error measure) works well and the entropy in the ensemble predicts well the reduction in error due to the ensemble. These measures are evaluated on a medical prediction problem and are shown to predict the performance of the ensemble well. We also show that the entropy measure of diversity has the added advantage that it seems to model the change in diversity with the size of the ensemble.

1. Introduction

Feature subset selection is an important issue in Machine Learning (Aha & Bankert, 1994; Bonzano, Cunningham & Smyth, 1997; Wettschereck, Aha, & Mohri, 1997). It is a difficult problem due to the potentially huge search space involved and because hill-climbing search techniques do not work so well because of an abundance of local maxima in the search space. Feature subset selection is important for the following reasons:

- **Build better predictors:** better quality predictors/classifiers can be build by removing irrelevant features – this is particularly true for lazy learning systems.
- **Economy of representation:** allow problems/phenomena to be represented as succinctly as possible.
- **Knowledge discovery:** discover what features are and are not influential in *weak theory* domains.

A different motivation for feature subset selection has emerged in recent years as illustrated in the research of Ho (1998a & 1998b) and Guerra-Salcedo and Whitley (1999a, 1999b). In their work feature subset selection is used as a mechanism for introducing diversity in ensembles of classifiers. Typically they work with datasets from weak theory domains where features have been oversupplied and there are irrelevant and redundant features in the representation.

In this paper we look at this approach to ensemble creation and propose entropy and cross entropy as measures of diversity and error that should be used in determining the constitution of the ensemble.

The paper starts with a review in section 2 of some existing research on ensembles of classifiers based on different feature subsets. We argue that diversity in the ensemble must be considered explicitly in putting together the ensemble. In section 3 we review the approach to diversity in regression ensembles where variance is the standard measure of diversity. We propose entropy as the appropriate measure in classification ensembles. In section 4 we present our algorithm for producing good quality feature subsets and in section 5 we show how the entropy measure of diversity provides a valuable insight into the operation of ensembles of classifiers in a medical application and helps determine the makeup of a very good ensemble.

2. Existing Research

Ho (1998b) introduces the idea of ensembles of Nearest Neighbour classifiers where the variety in the ensemble is generated by selecting different feature subsets for each ensemble. Since she generates these feature subsets randomly she refers to these different subsets as random subspaces. She points to the ability of ensembles of decision trees based on different feature subsets to improve on the accuracy of individual decision trees (Ho, 1998a). She advocates doing this also for *k*-Nearest Neighbour (*k*-NN) classifiers because of the simplicity and accuracy of the *k*-NN approach. She shows that an ensemble of k-NN classifiers based on random subsets improves on the accuracies of individual classifiers on a hand-written character recognition problem.

Guerra-Salcedo and Whitley (1999a, 1999b) have improved on Ho's approach by putting some effort into improving the quality of the ensemble members. They use a genetic algorithm (GA) based search process to produce the ensemble members and they show that this almost always improves on ensembles based on the random subspace process. The feature masks (subsets) that define each ensemble member are the product of GA search and should have higher accuracy than masks produced at random. The only situations where the random masks performed better that the masks produced by genetic search were on datasets with small numbers of features (19 features) (Guerra-Salcedo and Whitley 1999b).

Guerra-Salcedo and Whitley do not suggest any reasons why the random subspace method should out perform the genetic search method on data sets with small numbers of features. We suggest that this is explained by the analysis of diversity and accuracy presented in the next section. When genetic search is used to produce good feature subsets in small feature spaces the risk is that there will not be great diversity in the subsets produced. It is likely that the improvement in the quality of the individual classifiers is offset by the loss of diversity in the ensemble as a whole. This is not such a problem with datasets with large numbers of features (>35) because loss of diversity is less likely in such a huge search space. This diversity/quality issue will be discussed in detail in the next section. In concluding this paper we will argue that any work on ensembles should explicitly measure diversity and quality to ensure that the overall quality of the results of the ensemble is maximised.

3. Diversity

Krogh and Vedelsby (1995) have shown the following very important relationship between error and ambiguity (diversity) in regression ensembles

$$E = \overline{E} - \overline{A} \tag{1}$$

where E is the overall error of the ensemble over the input distribution, $\overline{E}$ is the average generalisation error or the ensemble components and $\overline{A}$ is the ensemble ambiguity averaged over the input distribution. $\overline{E}$ is a standard quadratic error estimation and $\overline{A}$ is an aggregation of individual ambiguities $\overline{a}(x)$, the ambiguity of a single ensemble member on a single input x:

$$\overline{a}(x) = \frac{1}{N}\sum_{n=1}^{N}(V^n(x) - \overline{V}(x))^2 \tag{2}$$

Thus the ambiguity is effectively the variance in the predictions coming from the ensemble members. This ambiguity can be tuned, for instance by overfitting neural networks, in order to maximise generalisation performance (Carney & Cunningham, 1998)

Because of the differences between ensembles of classifiers and regression ensembles it will not be straightforward to establish such a neat equality for classifiers. In particular the *winner-takes-all* nature of ensembles of classifiers radically changes the assessment of error. Once the ensemble produces the correct majority we *don't care* beyond that.

The obvious estimate of accuracy (or error) for a classifier is the proportion of a test set it classifies correctly.

$$e_i = P_{\hat{c}_i(x)=c(x)} \tag{3}$$

where $\hat{c}_i(x)$ is the category classifier i predicts for x and $c(x)$ is the correct category.

Then a possible measure of agreement (inversely related to ambiguity) is that used by Ho: using a test set of n fixed samples and assuming equal weights, the estimate of classifier agreement $\hat{s}_{i,j}$ can be written as:

$$\hat{s}_{i,j} = \frac{1}{n}\sum_{k=1}^{n} f(x_k) \tag{4}$$

where
$$f(x_k) = \begin{cases} 1 \text{ if } c_i(x_k) = c_j(x_k) \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

i.e. the measure of agreement is the proportion of test cases on which two classifiers agree. (Ho, 1998). Ho emphasises the importance of disagreement in ensemble members but does not directly evaluate its impact on the overall ensemble performance.

We will show in the evaluation section of this paper that a better measure of agreement (or ambiguity) for ensembles of classifiers is entropy. Tibshirani (1996) also suggests that entropy is a good measure of *dispersion* in bootstrap estimation in classification. So for a test set containing M cases in a classification problem where there are K categories a measure of ambiguity is:

$$\tilde{A} = \frac{1}{M}\sum_{x=1}^{M}\sum_{k=1}^{K} - P_k^x \log(P_k^x) \qquad (6)$$

where P_k^x is the frequency of the k^{th} class for sample x – the more dispersion or randomness in the predictions the more ambiguity. Associated with this entropy-based measure of diversity is a Conditional Entropy-based measure of error (loss function).

$$E_{CEnt} = \sum_{\hat{c}(x)\in K, c(x)\in K} P(\hat{c}(x), c(x)) \log P(\hat{c}(x) | c(x)) \qquad (7)$$

where $\hat{c}(x)$ is predicted category for sample x and $c(x)$ is the correct category. We will show in section 5 that if this measure is used as the loss (error) function the entropy measure of ambiguity in the ensemble better predicts the reduction in error due to the ensemble.

4. Producing Ensembles of Feature Masks

4.1. Random Masks

For a classification task with p possible input features there are 2^p possible subsets of this feature set and each subset can be represented as a feature mask of 1s and 0s. Masks of this type representing different feature subsets can easily be produced using a random number generator. These masks should score high on diversity because there has been no attempt to learn good quality feature sets. However, because of this, they cannot be expected to have very good scores for $\overline{E}$, the average error. Ho (1998b) has shown that ensembles of masks of this type can produce very good results – presumably because the lack of quality in the ensemble members is compensated for by the diversity.

4.2. Better Quality Masks

At the other end of the quality spectrum, Guerra-Salcedo & Whitley (1999a, 1999b) have used genetic algorithms (GA) to find high quality feature subsets. Since the GA search is, in Aha & Bankert terms, a *wrapper* process it is very computationally intensive because evaluating each state in the search space involves testing a classifier on a test set (Aha & Bankert, 1994). If this estimate of fitness is to be accurate then significant amounts of data must be used to build the classifier and test it. For this reason we use a simpler hill-climbing search technique that produces good quality masks but in reasonable time. The idea is to focus on managing diversity rather than ensemble member quality to provide overall ensemble quality. The algorithm for this is shown in Table 1.

Typically this algorithm will terminate after four cycles through the mask. At that stage there is no adjacent mask (i.e. a mask different in just one feature) that is better. Thus the masks produced are local maxima in the search space.

Table 1. Producing good quality feature masks using hill-climbing search.

We define $Acc(T_r, T_s, L)$ as the accuracy of a classifier on test T_s having been trained with T_r and using mask *L*. The accuracy is the proportion of T_s that is correctly classified.

1. Initialise mask *L* randomly as described in section 4.1.
2. Flag ← False
3. For each $l \in L$
   Produce L' from *L* by flipping *l*
   If $Acc(T_r, T_s, L') > Acc(T_r, T_s, L)$
      $L \leftarrow L'$
      Flag ← True
4. Repeat from 2 while Flag = True.

5. Evaluation

In this section we will assess this relationship between ambiguity and accuracy on some unpublished In-Vitro Fertilisation (IVF) data. The data consists of 1355 records describing IVF cycles of which 290 cycles have positive outcomes and 1065 have negative outcomes. In the representation of the data used here each data sample has 53 numeric input features. For the purpose of our evaluation 50 random masks were produced in the manner described in section 4.1. Then 50 better quality masks were produced in the manner described in section 4.2. To guide this search process 580 data samples were used including the 290 with positive outcomes – 330 are used in T_r and 250 in T_s.

This search process for producing the masks is very computationally expensive with the cost increasing with the square of the size of the data set used to guide the search. However if we skimp on the amount of data used, the masks will be biased towards the subset of data that does actually get used. Indeed it was clear during the course of the evaluation that the masks did overfit the training data, raising the question of overfitting in feature selection – a neglected research issue.

Then ensembles of size 5, 10, 20, 30, 40 and 50 were produced for the random masks and the better quality masks. These ensembles were tested using leave-one-out testing on the complete data set of 1355 samples. This means that the masks are being tested, in part, with the data used to produce them. This was done because of the small number of positive samples available and is reasonable because the objective is to show the ambiguity/accuracy relationship rather than produce a good estimate of generalisation error. Where possible, multiple different versions of the smaller ensembles were produced (i.e. 10 of size 5, 5 of size 10, 2 or size 20 and 2 of size 30). The results of this set of experiments are shown in Figure 1.

It can be seen that the random masks have an accuracy slightly inferior to the other masks averaging about 58.2% and 58.9% respectively (using a simple count of correct classifications as a measure of accuracy). For the various ensemble sizes there is very little difference in the diversity between the two scenarios. Thus the ensembles based on the better quality masks produce the best results with the ensemble of size 50 producing an accuracy in leave-one-out testing on the full data set of 64.5%. It is important to note

that this cannot be claimed as an estimate of generalisation accuracy for the system since the masks in use may overfit this data since some of the data was used in producing the masks.

An interesting aspect of the data shown in Figure 1 is that the measure of ensemble diversity used seems to capture the increase in diversity with ensemble size. As the benefit of increasing the ensemble size tails off around 30-40 members so does the increase in entropy. This would not be the case with the measure of diversity proposed by Ho (see section 3) for instance.

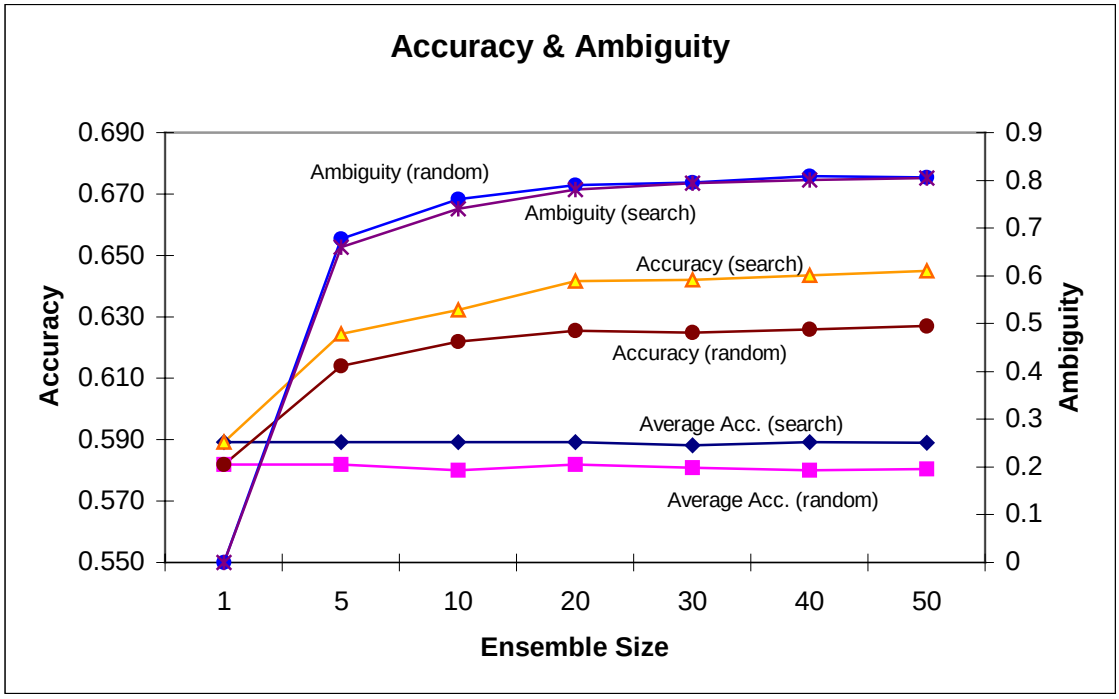
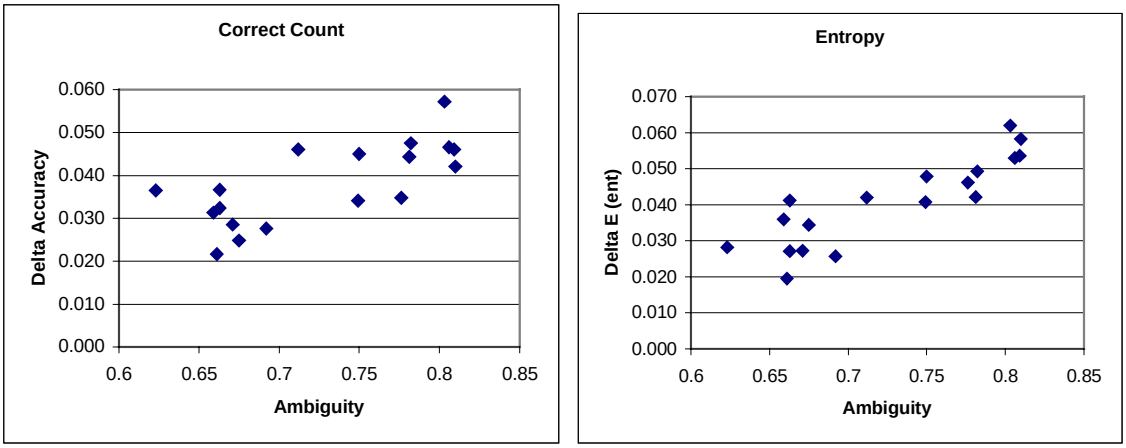


Figure 1. Measurements of accuracy and ambiguity of ensembles working on the IVF data.

This first experiment shows that if accuracy of ensemble members is increased without reducing ambiguity it will increase overall accuracy. In the next experiment we will show how the ambiguity of the ensemble predicts the reduction in error (increase in accuracy) due to the ensemble. These results are shown in Figure 2. In each graph the Y-axis shows the difference between the average error of the ensemble members and the ensemble error. In Figure 2(a) the error is a simple count of correct classifications; in (b) the conditional entropy is used as described in section 3. These graphs suggest that ambiguity as measured by entropy better predicts error reductions when error is measured as conditional entropy, i.e. the relationship in the graph on the right is clearer than that in the graph on the left. This is borne out by the correlation coefficient in both cases; the correlation coefficient for the relationship between change in correct count and entropy is 0.81 while that between change in error as measured by conditional entropy and ensemble entropy is 0.91.



(a) (b)

Figure 2. Plots of the relationship between the reduction in error due to an ensemble and the ambiguity of the ensemble. In (a) error is measured as a count of correct classifications; in (b) it is measured as conditional entropy.

Indeed it might be argued that, even without this useful relationship with ensemble ambiguity, conditional entropy is a particularly appropriate measure of error. Afterall, it does capture the importance of good accuracy spread across all categories. With the data presented here a good score based on count correct may conceal poor performance on the minority class.

Finally we show how the information provided by the use of entropy as a measure of diversity can inform the construction of a very good quality ensemble. The analysis shows that, in this domain, seeking good quality masks appears not to compromise ensemble diversity. We suggest that this is due to the large number of features in the domain (53). So it should be possible to increase the quality of the ensemble members without loss of diversity. 66 good quality masks were produced using the process described in section 4.2 and their accuracy was tested using leave-one-out testing on the full dataset of 1355 samples. Using this metric of quality the best 20 of these were chosen to form an ensemble. The accuracy of this ensemble measured using leave-one-out testing on the whole dataset was 66.9%, better than the average of 64.2% for the other two ensembles of size 20 and better than the 64.5% figure for best ensemble of size 50.

6. Conclusion

The main message in this paper is that any work with classification ensembles should explicitly measure diversity in the ensemble and use this measure to guide decisions on the constitution of the ensemble as shown in the last section.

We show that in the same way that variance is a good measure of diversity for regression problems entropy is a useful measure of diversity for classification ensembles. Then associated with entropy as a measure of diversity is conditional entropy as an appropriate error function.

As advocated by Ho and by Guerra-Salcedo and Whitley feature subsets are a useful mechanism for introducing diversity in an ensemble of *k*-NN classifiers. If the feature space under consideration is large (> 35) then there may be less risk of loss of diversity in searching for good quality ensemble members. In the future we propose to evaluate this

analysis on problems with smaller numbers of features where there may be a more clear-cut trade-off between ambiguity and quality of ensemble components.

Finally the quality of this ensemble of classifiers based on components with different feature subsets raises some questions about the issue of feature subset selection with which we opened this paper. The ensemble of classifiers has a better classification performance than any of its individual components. This brings into question the whole feature subset selection idea because it suggests that there is not one global feature set that provides a 'best' problem representation. Instead the ensemble exploits a variety of representations that may be combining locally in different parts of the problem space.

The next step is to evaluate these metrics on different classification datasets - However, leave-one-out testing on an ensemble of lazy learners is very computationally expensive. It will be particularly interesting to see if the entropy measure of diversity does in fact capture aspects of ensemble size and happens with this data set. For the future it will be interesting to tackle the problem of overfitting in the feature selection process.

Acknowledgements We would like to thank John Haslett for suggesting Entropy as a suitable measure of diversity and also for recommending Conditional Entropy as an associated loss function.

7. References

Aha, D. W., & Bankert, R. L. (1994). Feature selection for case-based classification of cloud types: An empirical comparison. In D. W. Aha (Ed.) *Case-Based Reasoning: Papers from the 1994 Workshop* (Technical Report WS-94-01). Menlo Park, CA: AAAI Press. (NCARAI TR: AIC-94-011).

Bonzano A., Cunningham P., Smyth B. (1997) Using introspective learning to improve retrieval in CBR: A case study in air traffic control, Case-Based Reasoning Research and Development, Proceedings of the *1997 International Conference on Case-Based Reasoning*, D.B. Leake and E. Plaza Eds., Springer Verlag, Lecture Notes in Artificial Intelligence, pp.291-302.

Carney, J., Cunningham, P., (1999) The NeuralBAG algorithm: optimizing generalization performance in bagged neural networks, in proceedings of *7th European Symposium on Artificial Neural Networks*, Bruges (Belgium), pp35-50 1999.

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.

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.

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.

Tibshirani, R., (1996) Bias, variance and prediction error for classification rules, University of Toronto, Department of Statistics Technical Report, November 1996 (also available at www-stat.stanford.edu/~tibs).

Wettschereck, D., Aha, D. W., & Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, pp273-314, Vol. 11, Nos. 1-5, 1977.