Case Representation Issues for Case-Based Reasoning from Ensemble Research

Pádraig Cunningham, Gabriele Zenobi

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

Abstract. Ensembles of classifiers will produce lower errors than the member classifiers if there is diversity in the ensemble. One means of producing this diversity in nearest neighbour classifiers is to base the member classifiers on different feature subsets. In this paper we show four examples where this is the case. This has implications for the practice of feature subset selection (an important issue in CBR and data-mining) because it shows that there is no *best* feature subset to represent a problem. We show that if diversity is emphasised in the development of the ensemble that the ensemble members appear to be local learners specializing in sub-domains of the problem space. The paper concludes with some proposals on how analysis of ensembles of local learners might provide insight on problem-space decomposition for hierarchical CBR.

1. Introduction

The idea of Case-Based Reasoning (CBR) has strong appeal because it is recognised that much of human expertise is experienced based and CBR is considered to capture this idea. Practitioners recognise that, while CBR does involve reuse of previous problem solving episodes in solving new problems, it is a limited imitation of the versatility of reuse that humans exhibit in problem solving. An important aspect of this shortcoming is the reliance of CBR on feature-vector representations of cases. This approach is firmly based in the symbolic AI paradigm (Newell & Simon, 1976) but it is recognised to have shortcomings. In 1997 an ECML workshop entitled "Case-Based Learning: Beyond Classification of Feature Vectors" (Wettschereck & Aha, 1997) was organised to advance the state of knowledge representation in CBR. However, it is fair to state that current CBR practice is still strongly dependent on cases represented as feature vectors.

In this paper we present an example of the shortcomings of a case representation based on a single feature-value vector. We present four examples of ensembles of *k*-Nearest Neighbour (*k*-NN) classifiers that outperform a single *k*-NN classifier based on the best single feature subset available. The ensemble incorporates several individual *k*-NN classifiers based on different feature subsets and *aggregates* their results. Based on our analysis of the ensembling process we will propose that its power derives from the aggregation of several *local* specialists.

While this illustrates the shortcomings of a solution based on a single global representation it is important to point out that approaches exist to address this within the CBR paradigm. The obvious solution is to have local weights that vary across the solution space and this approach has been explored in the lazy learning literature (Wettschereck, Aha, & Mohri, 1997). (Bonzano, Cunningham & Smyth, 1997) The difficulty with this approach is the problem of determining appropriate feature weights. It is well known that learning feature weights from examples is subject to overfitting. Kohavi, Langley & Yun,(1997) show that binary weights (i.e. feature selection) are often better that more fine grained weights due to the potential for overfitting. This problem is greatly exacerbated in local weighting because of the significant increase in the number of weights to be learned.

Another approach to this global representation problem within the CBR paradigm would be a hierarchical CBR approach where the problem space is partitioned with different sub-case-bases covering the different partitions. In fact, research on hierarchical CBR has focused on *problem* decomposition rather than *problem-space* decomposition (Smyth & Cunningham, 1992, Smyth et al., 2000). This is perhaps because of the difficulty in determining appropriate decompositions of the problem space. In the conclusion to this paper we propose that hierarchical CBR and problem-space decomposition may be the means to bring the performance of CBR systems up to that of ensembles. We outline how the analysis of ensemble solutions can provide guidance on problem-space decomposition.

The ensemble technique described in this paper achieves this decomposition implicitly with different members of the ensemble acting as specialists in different parts of the domain. While this approach is still a lazy learning technique with processing deferred to run time, it lacks the interpretability of CBR because of the large number of cases retrieved in the solution process. There is not a small number of cases that can be used as a starting point for adaptation or in explaining the answer.

In this paper we describe the operation of ensembles of nearest neighbour classifiers based on different feature subsets and show how such ensembles can perform better than any single classifier. We discuss the ramifications of this for CBR research. In the next section we describe ensembles in general and introduce the idea of ensembles based on different feature subsets. In section 3 we show how ensembles based on different feature subsets can outperform a classifier based on a single feature subset. Then the ramifications of this for CBR are discussed in section 4.

2. Ensembles

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 know now 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. Thus, a ready source of diversity is to train models on different subsets of the training data. This 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 (Ho,1998a;1998b), (Cunningham & Carney, 2000).

2.1. Ensembles and Lazy Learning

To an extent the strength of the ensemble idea is bad news for lazy learning research because, with the ensemble approach, several learners have to be consulted at run-time and lazy learners do all their work at run-time. With lazy learners there has been little or no offline processing of the training data to build a hypothesis or model – this work is deferred to run-time.

If the ensemble approach poses problems for lazy learning in general then it proposes fundamental problems for Case-Based Reasoning (CBR) in particular. The key idea in CBR is that a single case or a small number of cases are retrieved during problem solving and these cases are reused to produce a solution for the new problem. Thus, as a problem solving technique, CBR has the special advantage of interpretability. The motivation in ensemble research is quite different; instead of there being a single line of reasoning in the problem solving process there are several parallel lines of enquiry and the results of these parallel efforts are aggregated to form a single solution. Thus it is difficult to *explain* the output of an ensemble. For this reason it appears difficult to bring the advantages of the ensemble idea to CBR without losing the interpretability advantages of CBR.

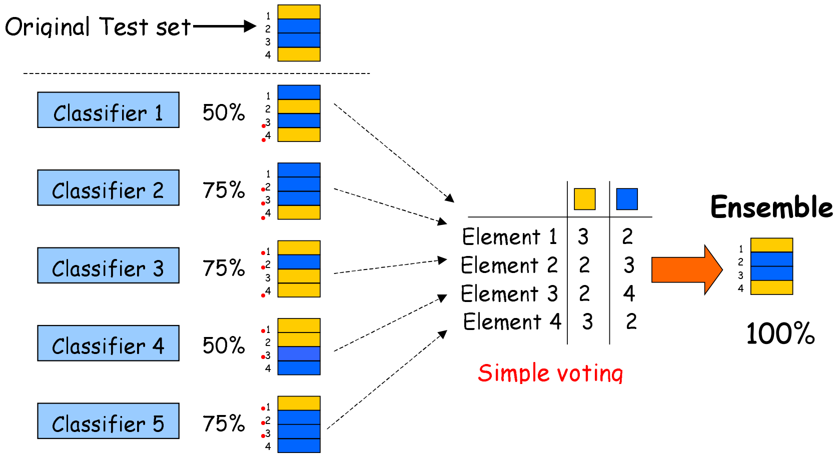
A further threat to the CBR idea arises from ensembles where the diversity in the ensemble is achieved by having different problem representations in the ensemble components. Research shows that ensembles of classifiers based on different feature subsets can produce better performance than a single classifier with a fixed feature set. (Cunningham & Carney, 2000), (van der Vaar & Heskes, 2000), (Ho,1998a;1998b) and (Guerra-Salcedo and Whitley,1999a; 1999b). The important point here is that this improved performance is not based on a single problem representation but on a variety of different representations. The ensemble might be considered to embody a distributed problem representation. Alternatively it might better be viewed as a committee of local problem solvers using a variety of local representations. This contrasts with the fixed representations that are common in case-based reasoning. In Richter's knowledge container terms (Richter, 1998), the vocabulary knowledge container is difficult to characterise in these ensembles based on different feature subsets.

2.2. The Importance of Diversity

Krogh & Vedelsby (1995) have shown that the reduction in error due to an ensemble is directly proportionate to the diversity or ambiguity in the predictions of the components of the ensemble as measured by variance. It is difficult to show such a direct relationship for classification tasks but it is clear that the *uplift* due to the ensemble depends on the diversity in the ensemble members. Cunningham and Carney (2000) propose that this diversity can be quantified using entropy.

Colloquially, we can say that; if the ensemble members are more likely on average to be right, and when they are wrong they are wrong at different points, then their decisions by majority voting are more likely to be right than that of individual members. But they must be more likely on average to be right and when they are wrong they must be wrong in different ways.

The example in Figure 1, if a little contrived, illustrates these principles in action. The test set is shown at the top of the figure and the five component classifiers have an average accuracy of 65% on this. The ensemble decision is by simple voting and the ensemble has an accuracy of 100% on the test data. If the four voting scenarios are examined it is clear that the tendency to be right coupled with the tendency to be wrong in different ways accounts for the improvement due to the ensemble.



	🟨	🟦
Element 1	3	2
Element 2	2	3
Element 3	2	4
Element 4	3	2

Fig. 1. An example showing how diversity in the ensemble leads to better classification.

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 data. This can be done systematically of by bootstrapping (sampling with replacement) different training sets from the training data. However, since this will not produce diversity in a stable (low variance) predictor such as *k*-NN (Breiman, 1996) it is of no use here.

Another popular technique is to use different feature subsets in the different classifiers – 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.

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:

$$a_i(x) = \left[V_i(x) - \overline{V}(x)\right]^2 \qquad \textbf{(1)}$$

where a_i is the ambiguity of the i^{th} classifier on example *x*, while V_i and $\overline{V}$ are, respectively the i^{th} classifier and the ensemble predictions.

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

$$a_i(k) = \begin{cases} 0 \text{ if } classV_i(k) = class\overline{V}(k) \\ 1 \text{otherwise.} \end{cases} \qquad \textbf{(2)}$$

where this time the classifier and ensemble outputs for the case labeled as *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:

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

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; instead we will use the 0/1 based approach as described in (2).

2.2. 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's 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 one another it reduces the other benefits of the feature selection strategy. It is quite intuitive 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.

Emphasizing Diversity

A very simple approach to selecting an ensemble of classifiers 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 Vedlsby (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 classifiers 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 provide some results for ensembles that are obtained with an algorithm based on a hill-climbing strategy that makes use of the concept of ambiguity (Zenobi & Cunningham, 2001). The results show that this technique outperforms the classic two-steps strategy for selecting ensembles via hill-climbing search. We give here only a brief description of it; a more detailed one can be found in the paper mentioned above.

In a classic hill climbing strategy 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 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 is the approach used to produce the results shown in Table 1.

The algorithm used to produce the results in Table 2 goes one step further; it considers instead every classifier in the context of the ensemble, and at each step accepts or rejects the flip depending on two parameters: the classifier error and the classifier ambiguity, using a variance measure as shown in equation (2). If the improvement of one of the two parameters leads to a "substantial" deterioration of the other, then the flip is rejected (see Figure 2). With "substantial" here we mean that a threshold is given for the highest acceptable deterioration. This technique allows us to avoid the selection of a set of good classifiers that make mistakes over the same subspace of the instances

```
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 {[( $E_i' \leq Thresh \times E_i$ )AND ($A_i' > A_i$ )] OR [( $E_i' < E_i$ )AND ($A_i' \geq Thresh \times A_i$ )]}
                $E_i = E_i'$ ; $A_i = A_i'$;   //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. The algorithm for generating ensembles while emphasising diversity in ensemble members.

Here we are not particularly interested in the difference in accuracy resulting from these two techniques. What is of interest is that the focus on diversity in the second technique appears to produce local learners.

3. Evaluation

The objective of this evaluation is to show that ensembles of *k*-NN classifiers can outperform the best single *k*-NN classifier available. We also show that if the ensemble members are forced to be diverse then very good ensemble accuracy can be achieved with ensemble members that have poor overall accuracy. These diverse ensemble members also prove to have fewer features that ensemble members selected without consideration for diversity.

We start with two data-sets, an InVitro Fertilisation (IVF) data-set presented in (Cunningham & Carney, 2000) and the Hepatitis data from the UCI repository. The IVF data has 53 features and Hepatitis has 19 features so it is not practicable to search through the full space of feature subsets to find the best mask. In each case 100 good masks are found using hill-climbing search as described above; see also (Cunningham & Carney, 2000). Six ensembles of size 20 are produced using these masks and the error figures are shown in Table 1. The ensembles outperform the best masks found.

Table 1. An error comparison of ensembles of *k*-NN classifiers compared with best single *k*-NN classifiers found using hill-climbing search.

Dataset	Average Mask	Best Mask Found	Average Ensemble	Best Ensemble
IVF	41.1%	38.3%	35.7%	35.5%
Hepatitis	20.7%	17.4%	16.8%	15.5%

In the next evaluation two data-sets are chosen with few enough features to perform an exhaustive search of the space of feature masks. These are the Pima and Abalone data-sets from the UCI repository and each has 8 input features. The Abalone data is converted to a classification format by partitioning the outcome variable into two classes (age 7&8 → Class 1; age 11,12&13 → Class 2). In this case ensembles are produced using the algorithm presented in Figure 2 that emphasizes diversity. These ensembles are compared using 5-fold cross validation with the best masks that have been found by exhaustive search. Using a one tailed paired *t*-test the ensemble is better than the best mask with 80% confidence for the Pima data and 90% confidence for the Abalone. 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 scoring of the masks in determining the best mask is done in the same fashion.

Table 2. An error comparison of ensembles of *k*-NN classifiers compared with best single *k*-NN classifiers.

Dataset	All Features	Best Mask	Ensemble	Average Ens. Mask
Pima	23.9%	23.8%	22.5%	30.7%
Abalone	18.1%	16.5%	15.9%	22.5%

The final part of the evaluation compares these ensembles with ensembles built considering error only (see Table 3). The overall error figures are slightly better when diversity is considered – however that is not the important point. 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 this small improvement in overall performance in the face of deterioration of the ensemble members is that the members are local specialists. This view is reinforced by the fact that the ensemble members produced using diversity have fewer features on average than the others (3.4 v's 4.6 for Pima and 3.7 v's 4.7 for Abalone). It seems reasonable that fewer features are required to discriminate in these local regions.

Table 3. A comparison of ensembles of *k*-NN classifiers trained using error and ambiguity and error only.

	Pima			Abalone		
	Average Error	Amb.	Ensemble Error	Average Error	Amb.	Ensemble Error
Error & Amb.	30.7%	21.5%	22.5%	22.5%	15.0%	15.9%
Error Only	26.9%	14.0%	23.9%	17.8%	5.6%	16.3%

3.1. Implications for CBR

From the results shown above we can make some observations. When we train an ensemble of classifiers we are able to obtain a better performance than any possible single classifier. This brings into question the practice of feature subset selection and illustrates some shortcomings of the feature vector representation of cases. Searching for *the* predictive feature subsets is questionable because a combination of feature subsets, maybe poorly predictive if taken singularly, can give a better prediction when aggregated in an ensemble. In other words, in a case-based reasoning system a single classifier (in this case, a single feature subset) might have an upper limit to its performance that can be raised by the use of an ensemble.

This result suggests the hypothesis that the search space is "adapted" by a single classifier (dropping some of the features) in order to fit better to its intrinsic metric (usually the Euclidean one). A single classifier uses a "global adaptation" of the space, while in the case of an ensemble we have a number of different "adaptations" that can be seen as making errors in different regions of the search space.

So in terms of the interpretability of the knowledge representation in the ensemble, at the extreme, there are two possibilities:

1. The ensemble embodies a distributed knowledge representation, similar to a neural network, and is not readily interpretable.
2. The ensemble is a set of local specialists and these specialists are locally interpretable.

Clearly, we are inclined to this second view. In our algorithm for generating diverse ensembles shown in Figure 2, by forcing the ensemble to be diverse the accuracy of the ensemble is increased but the average overall accuracy of the ensemble members is decreased. Coupled with this is the curious effect that the average number of features in the ensemble members is reduced compared to ensemble members that are optimized for accuracy only (see (Zenobi & Cunningham, 2001) for more details on this. We would suggest that this poor overall accuracy and small "specialized" feature set indicates that the ensemble members are local specialists.

This evaluation shows that an ensemble of lazy learners outperforms the best single lazy learner. The advantage of ensembling derives from aggregating diverse models. This advantage seems incompatible with CBR because if case-based learners are aggregated then the interpretability of CBR is lost.

In the future work section of this paper we propose a line of research whereby an ensemble of lazy learners would be used to discover a problem space decomposition that would allow for a hierarchical CBR solution that would have a performance comparable to that of the ensemble and better than a single CBR system.

4. Future Work and Conclusions

If, as suggested in section 4, the ensemble members are local specialists then the ensemble should be locally interpretable – the question is how to access this?

In this scenario local specialists are combining to ensure that elements in their locality are classified correctly. So one way to discover the problem decomposition implicit in the ensemble would be to cluster the data elements based on the classifiers that correctly classify them. For the problems presented in section 3 the classifiers that correctly classify the test data elements become *descriptors* of those data elements. Clustering the data according to these descriptors should produce clusters corresponding to the problem decomposition embodied in the ensemble.

These clusters should represent homogenous regions of the problem space where a single problem representation would work well. This suggests that a hierarchical CBR system that implements problem decomposition based on the problem decomposition implicit in the ensemble should perform better than a single CBR system operating over the whole problem domain.

4.1. Conclusion

In this paper we show how an ensemble of lazy learners based on different feature subset can perform better than a single lazy learner based on the best feature subset available. This brings into question the practice of searching for a single best feature subset, an issue that has received a lot of attention in CBR and in data-mining. This may indicate a fundamental shortcoming of symbolic representations or, more likely, it simply shows that a single representation across the whole problem space is not adequate

We show that if diversity is emphasized in the ensemble that the ensemble members act as local learners specializing in sub-regions of the problem space. We propose that an analysis of the performance of the ensemble may provide an insight into how the problem space should be partitioned to develop a hierarchical CBR system that implements problem space decomposition.

5. References

Aha, D.W. (1998) Feature weighting for lazy learning algorithms. In: H.Liu and H. Motoda (Eds.) *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Norwell MA: Kluwer.

Bonzano, A., Cunningham, P., Smyth, B., (1997) Using introspective learning to improve retrieval in CBR: A case study in air traffic control *International Conference on Case-Based Reasoning,* Providence, Lecture Notes in Computer Science, SpringerVerlag, E. Plaza & D. Leake (eds), pp291-302.

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

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.

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.

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.

Kohavi, P. Langley, Y. Yun, (1997) The Utility of Feature Weighting in NearestNeighbor Algorithms, *European Conference on Machine Learning, ECML'97*, Prague, Czech Republic, 1997, poster.

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.

Newell, A., & Simon, H.A., (1976) Computer Science as Empirical Enquiry: Symbols and Search. *Communications of ACM*, 19(3), 1976, pp.113-126.

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.

Richter, M. M. (1998). Introduction (to Case-Based Reasoning). in *Case-based reasoning technology: from foundations to applications*, Lenz, M., Bartsch-Spörl, B., Burkhard, H.-D. & Wess, S. (eds.) (1998). Springer-Verlag, LNAI 1400, pp1-16.

Smyth B., Cunningham P., (1992) Déjà Vu: A Hierarchical Case-Based Reasoning System for Software Design, in Proceedings of *European Conference on Artificial Intelligence*, ed. Bernd Neumann, John Wiley, pp587-589, Vienna Austria, August 1992.

Smyth, B., Keane, M., & Cunningham, P., (2000) Hierarchical Case-Based Reasoning: Integrating Case-Based and Decompositional Problem-Solving Techniques for Plant-Control Software Design, *to appear in IEEE Transactions on Knowledge and Data Engineering.*

van de Laar, P., Heskes, T., (2000) Input selection based on an ensemble, *Neurocomputing*, 34:227-238.

Wettschereck, D., & Aha, D. W. (Eds.) (1997). ECML-97 MLNet Workshop Notes: *Case-Based Learning: Beyond Classification of Feature Vectors* (Technical Report AIC-97-005). Washington, DC: Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence.

Wettschereck, D., Aha, D. W., & Mohri, T. (1997). A review and comparative evaluation of feature weighting methods for lazy learning algorithms. *Artificial Intelligence Review, 11*, 273-314.

Zenobi, G., & Cunningham, P., (2001) Using ambiguity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error, *submitted to ECML 2001.*