An Approach to Aggregating Ensembles of Lazy Learners that Supports Explanation[1]

Gabriele Zenobi, Pádraig Cunningham

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

Abstract: Ensemble research has shown that the aggregated output of an ensemble of predictors can be more accurate than a single predictor. This is true also for lazy learning systems like Case-Based Reasoning (CBR) and *k*-Nearest-Neighbour. Aggregation is normally achieved by voting in classification tasks and by averaging in regression tasks. For CBR, this increased accuracy comes at the cost of interpretability however. If we consider the use of retrieved cases for explanation to be one of the advantages of CBR then this is lost in an ensemble. This is because a large number of cases will have been retrieved by the ensemble members. In this paper we present a new technique for aggregation that obtains excellent results and identifies a small number of cases for use in explanation. This new approach might be viewed as a transformation process whereby cases are transformed from their feature based representation to a representation based on the predictions of ensemble members. This new representation produces very accurate predictions and allows a small number of similar neighbours to be identified.

1 Introduction

A major development in Machine Learning (ML) research in recent years is the realisation that ensembles of models can offer significant improvements in accuracy over single models. To define an ensemble we need two elements: a set of properly trained classifiers and an aggregation mechanism that composes the single predictions into an overall outcome. Typically, the aggregation process will be a simple average or a simple majority vote over the output of the ensembles, e.g. (Breiman, 1996). However, it may also be a complex linear or non-linear combination of the component predictions, e.g. (Jacobs et al., 1991, Heskes, 1998).

Whatever the aggregation process, it has implications for interpretability if the ensemble is composed of lazy learners (Cunningham & Zenobi, 2001). This is very important for domains such as medical decision support where interpretability plays a fundamental role. For instance Ong et al. (1997) and Armengol et al. (2001) describe CBR systems for medical decision support where the use of retrieved cases in explanation plays a central role (see also (Leake, 1996) on explanation in CBR).

[1] This research was carried out as part of the MediLink project funded under the PRTLI programme of the Irish Higher Education Authority.

CBR allows for the use of the retrieved cases in explanation as follows:

"The system predicts that the outcome will be X *because that was the outcome in case* C1 *that differed from the current case only in the value of feature* F *which was* f2 *instead of* f1.

In addition the outcome in C2 was also X *..."*

Explanation in these terms (i.e. expressed in the vocabulary of the case features) will not always be adequate, but in some situations such as in medical decision support it can be quite useful. However if the prediction is coming from an ensemble of CBR systems rather than a single system there is no longer a small number of cases to use for explanation.

So there appears to be a fundamental incompatibility between the ensemble idea and the interpretability of CBR. By definition, the ensemble is an order of magnitude more complex than a basic CBR system with an extra layer of processing (i.e. aggregation) between the cases and the proposed solution. In (Zenobi & Cunningham, 2001) we have argued that the effectiveness of ensembles stems in part from the ensemble performing an implicit decomposition of the problem space with ensemble members *specializing* in local regions of the space. Presumably an explanation of the output of the ensemble should also reflect the way the ensemble has modeled the problem space.

In this paper we present a new approach to the ensemble aggregation process that obtains excellent results and identifies a small number of cases for use in explanation. This new approach might be viewed as a representation transformation process whereby cases are transformed from their feature-based representation to a representation based on the predictions of ensemble members (see section 3). If this representation is used for prediction using a simple nearest neighbour approach (we call this *Meta k*NN) it has a generalization accuracy comparable to that of the ensemble. We argue that this is because it accesses the model of the problem domain that is implicit in the ensemble. This view is supported by the fact that the *Meta k*NN shows very high fidelity to the ensemble predictions. The evaluation in section 4 also shows that this *Meta k*NN classifier produces very accurate predictions and allows a small number of similar neighbours to be identified. But first, the process of aggregating a set of case-based classifiers into an ensemble is described in the next section.

2 Ensembles of *k*-Nearest Neighbour Classifiers

It is well known that ensembles of predictors can improve on the performance of a single predictor (Hansen & Salmon, 1990; Krogh & Vedelsby, 1995; Breiman, 1996). This improvement depends on the members of the ensemble being diverse; a characteristic that arises naturally with decision trees or neural networks trained using different data sets. Indeed the ensemble has the added advantage of overcoming this instability problem. *k*-Nearest-Neighbour (*k*-NN) classifiers do not have this instability so producing an ensemble that will show an *uplift* requires another approach. The most common way to do this is to base the ensemble members on different feature subsets (Ho, 1998; Guerra-Salcedo & Whitley, 1999a, 1999b).

Again, it has been shown that the improvement due to the ensemble depends on the diversity of the members (Ricci & Aha, 1998;Cunningham & Carney, 2000; Zenobi & Cunningham, 2001).

Figure 1 shows how such an ensemble of *k*-NN classifiers would operate. Assume that the ensemble members are based on different feature subsets and these subsets have been chosen to maximize diversity and minimize error (see section 2.1). In this example there are *m* classifiers and *k* is set to 3. The task is binary classification with black corresponding to 1. The first classifier retrieves 2 black and one white example. By simple voting this will predict black (1) as the output; alternatively a fuzzy or probabilistic prediction might be produced as follows:

$$p(c_j|x) = \frac{\sum_{k \in \mathbf{K}} 1(k_c = c_j) \cdot 1/d(k,x)}{\sum_{k \in \mathbf{K}} 1/d(k,x)} \quad (1)$$

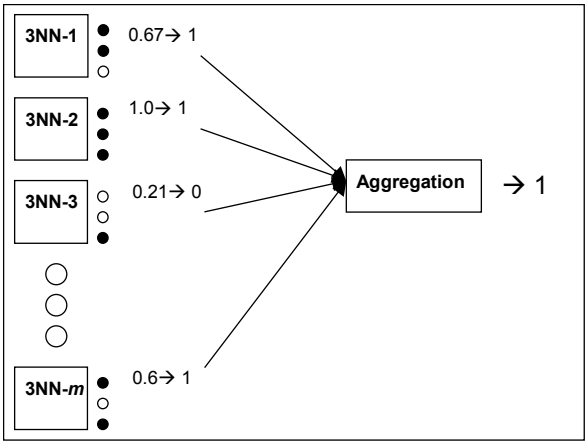

Fig.1. The aggregation process in an ensemble of *k*-Nearest Neighbour Classifiers.

This is the probability of example x having class c_j, where **K** is the set of k nearest neighbours of x, k_c, is the class of k, $d()$ is the distance function and $1()$ returns 1 iff its argument is true (Wettschereck et al. 1997). A probabilistic prediction of 0.67 would translate to a prediction of 1 that would be passed to the aggregation process. Alternatively the prediction of 0.67 could be passed for aggregation. The relevance of this is that the aggregation process will perform averaging for continuous predictions and voting for binary predictions. These alternatives generalize to the multi-class situation in a straightforward manner. In the evaluations presented in this paper it is the continuous (probabilistic) value that is passed to the aggregation step. Then the prediction of the ensemble is a weighted sum of the component predictions:

$$p_E(c_j|x) = \sum_{i=1}^{m} w_i p_i(c_j|x) \quad (2)$$

where p_i is the probabilistic prediction of the i^{th} member – in our implementation the weights are equal and sum to 1.
The key point in presenting this example is to show that a large number of cases contribute to the prediction of the ensemble. For example, in an ensemble of 25 classifiers each retrieving 5 neighbours, 125 neighbours will be retrieved. Clearly there will be duplicates in this set but with the prediction from each classifier being made on the basis of a different feature subset, it is very likely that the nearest neighbours for one particular classifier will not be the same as for another. When the neighbours from all members are put together in a "pool" it is not clear which ones are the most representative. We might rank based on frequency of occurrence but that is unlikely to be a complete solution. In section 3 we present our *Meta k*-NN solution to this problem but first we complete the discussion on ensembles with an account of how to train ensembles of *k*-NN classifiers in order to maximise diversity.

2.1 Training with Diversity

It is well known that the potential for an ensemble to be more accurate than its constituent members depends on there being diversity in the ensemble. If all ensemble members agree, there is no *uplift* due to the ensemble; instead it is important for the ensemble members to be right (and wrong) in different areas of the problem space (Zenobi & Cunningham, 2001).

For classification the most commonly used error measure is a simple 0/1 loss function, so a measure of diversity (ambiguity) on a single prediction is:

$$a_i(x_j) = \begin{cases} 0 \text{ if } \arg\max_{c_j}[p_j(c_j \mid x)] = \arg\max_{c_j}[p_E(c_j \mid x)] \\ 1 \text{ otherwise.} \end{cases} \quad \textbf{(1)}$$

where $a_i(x_j)$ is the ambiguity of the i^{th} classifier on example x_j and the two *arg max* functions return respectively the i^{th} classifier and the ensemble predicted class. Thus the contribution to diversity of an ensemble member i as measured on a set of N examples is:

$$A_i = \frac{1}{N}\sum_{j=1}^{N} a_i(x_j) \quad \textbf{(2)}$$

So an ensemble trained to minimize the error of individual members while maximizing this contribution to diversity will be very effective. An algorithm to do this for ensembles of *k*-NN classifiers based on different feature subsets is described in (Zenobi & Cunningham, 2001). The details of this algorithm will not be repeated here but the basic principles are as follows. Each ensemble member is defined by a feature mask identifying the features that are 'turned on' in that member. The training of the ensemble involves searching through the space of all feature masks to identify a set of masks that maximizes the diversity and minimizes the error of individual members. This search involves flipping bits in the masks and testing to see if they produce an improvement in error or diversity. Since there is clearly a tradeoff

between error and diversity a threshold is used within which small deteriorations in one measure are tolerated for the sake of significant improvements in the other (see (Zenobi & Cunningham, 2001) for details).

This process produces an ensemble of *k*-NN classifiers based on different feature subsets that is very accurate but difficult to interpret due to the potentially large number of cases involved in generating a solution.

In the evaluation presented in section 4 two different approaches to producing ensembles of *k*-NN classifiers are evaluated. These are called:

- *AmbHC:* the method described here that uses diversity (i.e. ambiguity+hill-climb)
- *HC:* search for feature masks is based on error only, does not consider diversity.

In both these scenarios the aggregation for the ensemble is done using the weighted sum described in equation 2 (ensemble members are assigned equal weights).

3 The Meta *k*-Nearest-Neighbour Aggregation Technique

Consider a database consisting of a set of *n* cases, each one described by *f* features (see Table 1). For simplicity suppose all the features are normailsed numerical and that the outcome is a simple binary classification mapped to the classes 0 and 1.

Table 1. A sample data set of *n* cases each described by *f* features.

	Feature 1	**Feature 2**	**…**	**Feature *f***	**CLASS**
Case 1	0.23	0.16	…	0.98	0
Case 2	0.14	0.56	…	0.32	1
Case 3	0.45	0.16	…	0.42	0
…	…	…	…	…	…
Case *n*	0.56	0.18	…	0.0	1

Suppose we train an ensemble of *m k*-NN classifiers differing on the feature subset chosen as described in section 2.1. For the training data, each classifier will return a class prediction (in the form of a probability between 0 and 1). It is then possible to associate a new $n \times m$ matrix with this ensemble, where in the position *(i,j)* is stored the prediction given by the classifier *j* for the case *i*. In other words each case is described by a new set of features representing how the ensemble (through each one of its classifiers) *sees* the case. An example of such a matrix is shown in Table 2. The arrows indicate what would have been the final prediction if the classifier were used on its own.

This new matrix is a transformation of the data that in some sense reflects how the ensemble has modelled the problem domain. It also suggests a new two-stage process of classification. In the first stage a target example is presented to the ensemble as before. In the second stage the outputs of the ensemble members is used as a representation of the case in a *Meta k*-NN classification process. The case-base for the *Meta k*-NN process is the transformed data shown in Table 2.

Table 2. A transformation of the data shown in Table 1 based on the outputs of the *m* classifiers in the ensemble.

	Classifier 1	**Classifier 2**	**…**	**Classifier *m***	**CLASS**
Case 1	0.95 →1	0.08 →0	…	0.21 →0	0
Case 2	0.67 →1	1.0 →1	…	0.0 →0	1
Case 3	0.21 →0	0.19 →0	…	0.69 →1	0
…	…	…	…	…	…
Case *n*	0.61 →1	0.32 →0	…	0.15 →0	1

This *Meta k*-NN classifier has excellent accuracy – equivalent to that of the ensemble and has the added advantage that a small number of cases are identified for use in explanation.

4 Evaluation and Discussion

In this section we present an experimental study of the aggregation technique we have described in section 3. This evaluation shows two things:

i. Using the *Meta k*-NN aggregation technique we obtain performance (accuracy) that is comparable to that obtained with the standard weighted average technique. This is shown by a comparison of the generalization errors of both the techniques.
ii. The *Meta k*-NN technique, which is a single classifier working with a transformed representation produced by the ensemble, models the problem domain in a way that is very similar to the ensemble on which it is based. This is shown by measuring the fidelity of the predictions from the *Meta* k-NN technique to the predictions given by standard aggregation technique.

Since these evaluations are very computationally intensive we present results on only four datasets, three from the UCI repository (Pima Indians, Heart Disease, Cylinder) and the Warfarin data-set described in (Byrne et al., 2000).

We have focused for simplicity on binary classification problems (in the case of Warfarin we have turned it into a 2-class task). The *Meta k*-NN aggregation technique is easily generalized to the case of *n*-classes. We have also considered datasets that do not have a skewed class distribution, as simple 0/1 error measures are questionable for datasets with very unbalanced class distributions.

In the following set of four figures (Fig. 2, 3, 4 and 5) we show the first of the two studies mentioned above; each figure refers to a different data set. For a complete comparison we have applied to each dataset both the *HC* and *AmbHC* training algorithms described in (Zenob i& Cunningham, 2001). Using these we have trained ensembles of 25 *k*-NN classifiers (k =5). The generalization error of each ensemble was determined using 5-fold cross validation. This entire process was repeated 2 or 3 times and the results averaged since the hill-climbing strategy is quite sensitive to the initial condition.

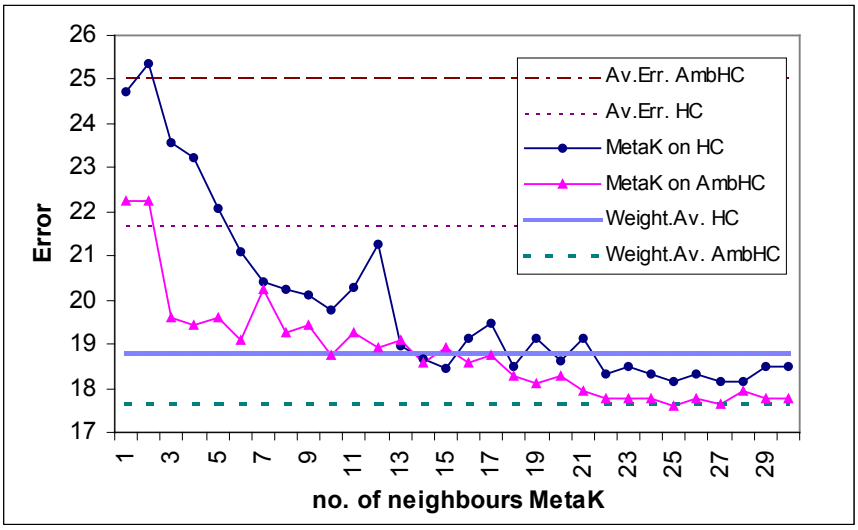

Fig. 2. Results for Heart data

The diagrams show the generalization error of the *Meta* k-NN aggregation strategy (both for ensembles trained with *HC* and *AmbHC*) plotted against the number of retrieved neighbours k_M. It is important to distinguish k_M from k, which is the number of retrieved neighbours for any single classifier in the ensemble. It is worth noting that the choice of this k_M is completely unrelated to the choice of k.

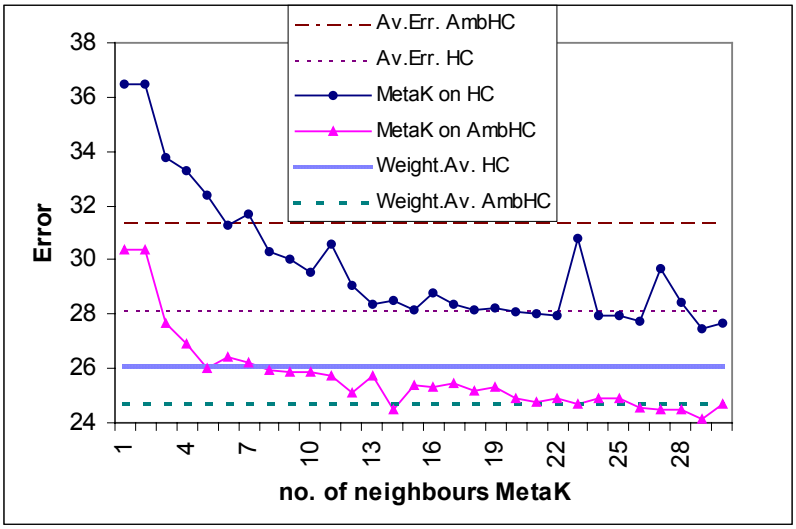

Fig. 3. Results for Pima data

To facilitate comparisons we also show the generalization error of the standard weighted average technique (both for ensembles trained with *HC* and *AmbHC*) and the average generalization of the component classifiers in the ensembles. These four figures appear as horizontal lines as they obviously do not depend on the value of k_M.

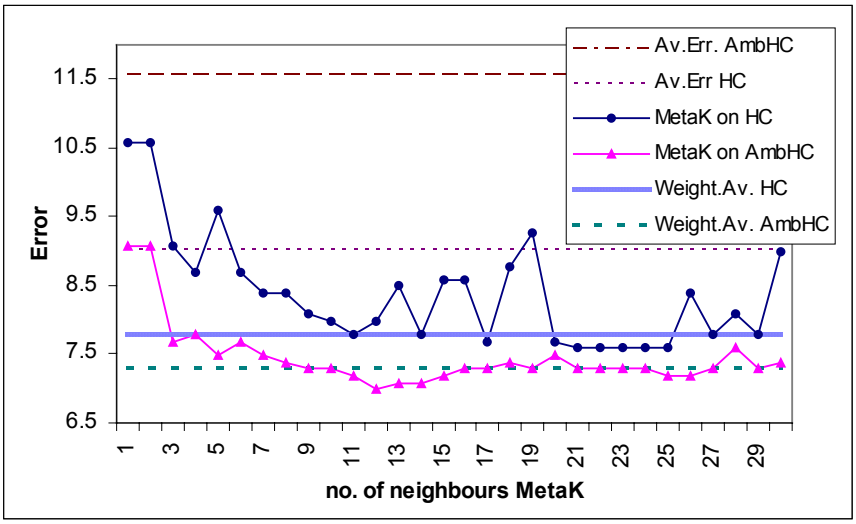

Fig. 4. Results for Warfarin data

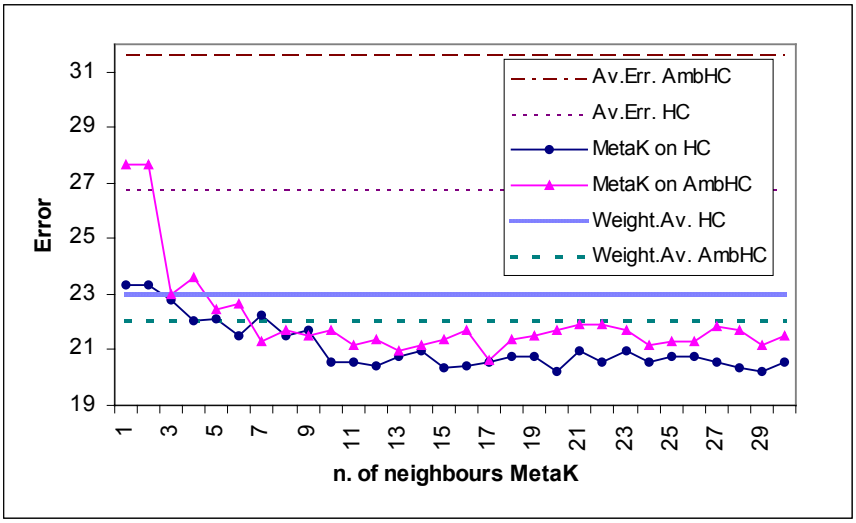

Fig. 5. Results for Cylinder data

From the figures shown above we can make a few observations. First, there is a confirmation of the fact that ensembles trained with *AmbHC* and using the standard aggregation approach have a better generalization error that those trained with a simple *HC*, even though the average error of the component classifiers are considerably worse. This is no surprise, and simply shows once again that diversity plays a crucial role in the ensemble performance.

Second, the *Meta* k-NN aggregation technique performed on ensembles trained with *AmbHC* scores comparable or better results to the weighted average aggregation technique for larger values of k_M. As a general trend we observe that already for values of k_M between 5 and 7 the result obtained by *Meta* k-NN is comparable to a weighted average on ensembles trained with *HC* (i.e. a score about between 1% and 2.5% worse than the one for *AmbHC*), with the great advantage of retrieving a small set of cases for explanation. When we increase the value of k_M up to 25 the *Meta* k-NN technique outperforms (except in the Heart dataset) ensembles trained with *AmbHC* and using the standard aggregation technique. In this case the set of retrieved neighbours is large (but still smaller than that retrieved by a classic aggregation technique) and has the important capability of giving a coherent ranking to the cases retrieved.

Third, the *Meta* k-NN aggregation technique performed on ensembles trained with simple *HC* scores generally worse generalization errors than the one performed on ensembles trained with *AmbHC*. A possible explanation for this phenomenon can be the fact that classifiers trained with *HC* have a lower diversity, so they carry less "rich" information about the problem domain, than the ones trained with *AmbHC*; when the *Meta* k-NN case-base is created it is possible that the columns (classifiers' predictions) show a higher dependence in the case of *HC*. The only exception to this is the Cylinder dataset where the *AmbHC* approach has no clear advantage over the *HC* technique. This is probably due to the large number of features in this dataset (38) and the consequent 'natural' diversity that exists even in ensemble members trained using *HC*.

Figure 6 shows the second study mentioned at the beginning of this section. It compares the fidelity of the *Meta k*-NN classifiers with the corresponding *AmbHC* ensemble. We do not consider the *HC* ensemble because it (and the corresponding *Meta k*-NN classifier) have poorer accuracy.

We have plotted for all four datasets the figures for fidelity between the predictions given by the *Meta k*-NN and standard aggregation technique. This is calculated as a binary error (0 if the two predictions match, 1 if they don't) and is plotted against increasing values of k_M.

From this figure it is clear that the fidelity is high. After k_M=5 already all the datasets (except Cylinder) score a fidelity over 95% (error less than 5%) and for two of the datasets the fidelity goes up to 98% and more as k_M increases. We can reasonably argue that the problem domain decomposition in the case of the two different aggregation strategies is nearly equivalent.

5. Conclusions and Future Research

Ensembles have had a big impact on Machine Learning research in recent years because they bring significant improvements in accuracy and stability. Another development in ML research is the emphasis on interpretability explanation. This is probably due to the increased interest in Data Mining where the emphasis is as much on insight as prediction. Because ensembles introduce an extra layer of complexity they make explanation much more difficult. In this paper we have presented a technique that reconciles these two things – at least for lazy learning systems.

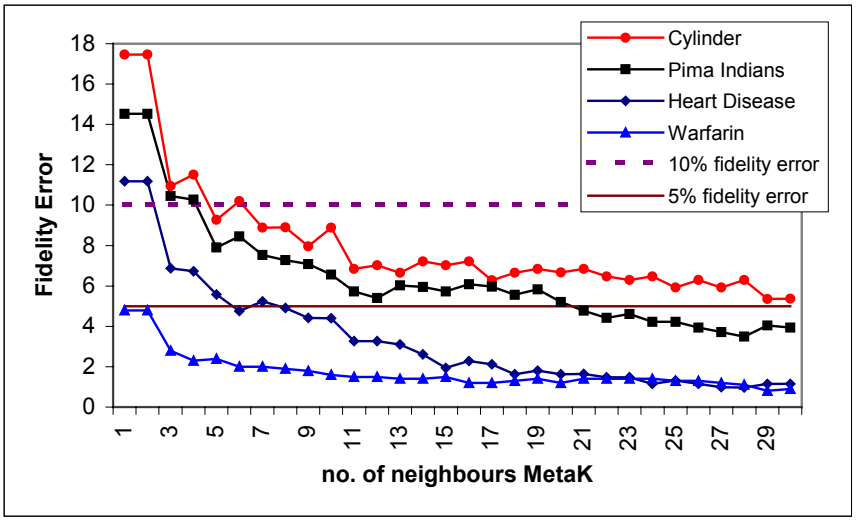

Fig. 6. Fidelity error of the *Meta k*-NN classifiers to the ensemble on which they are based. The 10% and 5% error lines are shown for comparison.

We have introduced a new technique, *Meta* k-NN, that performs the aggregation step when using an ensemble. This technique obtains good results, comparable to the standard averaging approach to aggregation in terms of generalization error, and allows us to identify a small set of cases for use in explanation. Even though large vales of k_M are required to provide good accuracy this is not a problem as the set of cases are ranked and the top ranking cases can be used for explanation.

In conclusion, we have introduced a new aggregation process that might be used in two modes:

i. Use *Meta k*-NN to produce the prediction *and* to identify cases for use in explanation. This would be appropriate when *Meta k*-NN showed to have a generalization accuracy equal to that of the standard aggregation technique.

ii. Use the standard aggregation technique to produce predictions and use *Meta k*-NN to identify cases for explanation – the high fidelity would allow for this. This would be appropriate when the accuracy of *Meta k*-NN was poorer than the standard aggregation.

5.1. Future Work

Since the key benefit we claim for this technique is its ability to select cases for use in explanation we need to evaluate the usefulness of the cases retrieved. We have access to domain experts in the Warfarin domain and in other medical domains and we will perform a study where these experts will rate the relevance of the retrieved cases.

The accuracy of *Meta k*-NN might further be improved by performing feature subset selection. Some of the features (i.e. ensemble member predictions) are probably more informative than others and deleting some features may improve performance.

References

Armengol, E., Palaudàries, A., Plaza, E., (2001). Individual Prognosis of Diabetes Long-term Risks: A CBR Approach. *Methods of Information in Medicine. Special issue on prognostic models in Medicine*. vol. 40, pp. 46-51.

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

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, in Proceedings of 4th International Conference on Case-Based Reasoning eds D. W. Aha, I. Watson, LNAI 2080, pp146-157, 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.

Heskes, T.M. (1998). Selecting weighting factors in logarithmic opinion pools. *Advances in Neural Information Processing Systems, 10*, 266-272.

Ho, T.K., (1998) 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, 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.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., & Hinton, G.E., (1991) Adaptive mixtures of local experts, *Neural Computation*, 3, 79-98.

Leake, D., B., (1996) CBR in Context: The Present and Future, in Leake, D.B. (ed) Case-Based Reasoning: Experiences, Lessons and Future Directions, pp3-30, MIT Press.

Ong, L.S., Sheperd, B., Tong, L.C., Seow-Choen, F., Ho, Y.H., Tong, L.C., Ho Y.S, Tan, K. (1997) The Colorectal Cancer Recurrence Support (CARES) System. Artificial Intelligence in Medicine 11(3): 175-188.

Ricci, F., & Aha, D. W. (1998). Error-correcting output codes for local learners. Proceedings of the Tenth European Conference on Machine Learning (280-291). Chemnitz, Germany: Springer.

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

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, 11*, 273-314.

Zenobi, G., Cunningham, P., (2001) Using Diversity in Preparing Ensembles of Classifiers Based on Different Feature Subsets to Minimize Generalization Error, 12th European Conference on Machine Learning (ECML 2001), eds L. De Raedt & P. Flach, LNAI 2167, pp576-587, Springer Verlag.