Explaining Predictions from a Neural Network Ensemble One at a Time

Robert Wall, Pádraig Cunningham, and Paul Walsh

Department of Computer Science, Trinity College Dublin

Abstract. This paper introduces a new method for explaining the predictions of ensembles of neural networks on a case by case basis. The approach of explaining individual examples differs from much of the current research which focuses on producing a global model of the phenomenon under investigation. Explaining individual results is accomplished by modelling each of the networks as a rule-set and computing the resulting coverage statistics for each rule given the data used to train the network. This coverage information is then used to choose the rule or rules that best describe the example under investigation. This approach is based on the premise that ensembles perform an implicit problem space decomposition with ensemble members specialising in different regions of the problem space. Thus explaining an ensemble involves explaining the ensemble members that best *fit* the example.

1 Introduction

Neural networks have been shown to be excellent predictors. In many cases their prediction accuracy exceeds that of more traditional machine learning methods. They are, however, unstable. This means that although two networks may be trained to approximate the same function, the response of both neural networks to the same input may be very different. Ensembles of networks have been used to counteract this problem. An ensemble comprises a group of networks each trained to approximate the same function. The results of executing each of these networks is then combined using a method such as simple averaging [3] in the case of regression problems, or voting in the case of classification problems. Ensembles used in this way show great promise not only in increasing the stability but also the accuracy of neural networks. The more diverse the members of the ensemble, the greater the increase in accuracy of the ensemble over the average accuracy of the individual members [6].

A further problem with neural networks is their 'black box' like nature. Users are not able to interpret the complex hyperplanes that are used internally by the network to partition the input space. A neural network may prove to be a better predictor for a particular task than alternative interpretable approaches but it is a black box. Therefore, substantial research has been done on the problem of translating a neural network from its original state into alternative more understandable forms. However, despite the obvious advantages of ensembles, much

T. Elomaa et al. (Eds.): PKDD, LNAI 2431, pp. 449–460, 2002.

less work has been done on the problem of translating ensembles of networks into more understandable forms.

Zenobi & Cunningham [14] argue that the effectiveness of ensembles stems in part from the ensemble performing an implicit decomposition of the problem space. This has two consequences for explaining ensembles. First it implies that a comprehensible model of the ensemble may be considerably more complex than an individual network. But more importantly, it means that parts of the ensemble will be irrelevant in explaining some examples.

Due to the increased complexity of ensembles, the objective of producing a global explanation of the behaviour of the ensemble is very difficult to achieve. So the goal of our research is focused on explaining specific predictions - a goal that is achievable for ensembles. Whereas, in this paper we concentrate on explaining ensembles of neural networks, our approach can be applied to any ensemble where outputs of an ensemble member can be explained by rules. This local, rather than global, approach to explanation is further elaborated in the next section. A brief introduction to the types of neural networks investigated is given in section 3.2. The behaviour of individual networks is modelled using rules derived from a decision tree that is built to model the outputs of an individual neural network, this is discussed in section 3.3. A method for selecting the most predictive of these rules for any given case is then presented in section 3.4. Also included in section 3.5 are some comments on how different policies may be used in different circumstances depending on the user of the system. Finally, section 4 includes an evaluation of the results with comments from an independent expert in the area of study.

2 Explanation

Explanation is important in Machine Learning for three reasons:

- to provide insight into the phenomenon under investigation
- to explain predictions and thus give users confidence
- to help identify areas of the problem space with poor coverage, allowing a domain expert to introduce extra-examples into the training set to correct poor rules

The first of these objectives is 'Knowledge Discovery' and can be achieved by producing a global model of the phenomenon. This global model might be a decision tree or a set of rules. Since Machine Learning techniques are normally used in *weak theory* domains it is difficult to imagine a scenario where such a global model would *not* be of interest. The second objective is more modest but we argue is adequate in a variety of scenarios. In the next two subsections we discuss why producing global explanations of ensembles is problematic and consider situations where local (i.e. example oriented) explanation is adequate.

2.1 Explaining Neural Networks

Many domains could benefit greatly from the prediction accuracy that neural networks have been shown to possess. However, becuase of problems with the black-box nature of neural networks(particularly in domains such as medical decision support), there is a reluctance to use neural networks. Capturing this prediction accuracy in a comprehensible format is behind the decision to generate rules based on neural networks in this research.

Most of the work on explaining neural networks has focused on extracting rules that explain them; a review of this work is available in [13]; a more in depth discussion of specific methods is available in [1].

The research on rule extraction can be separated into two approaches, direct decomposition and black box approaches. In a direct decomposition approach interpretable structures (typically trees or rules) are derived from an analysis of the internal structure of the network. With black box approaches, the internals of the network are not considered, instead the input/output behaviour of the network is analysed (see section 3.3). Clearly, the first set of techniques is architecture-specific while the black-box approaches should work for all architectures. The big issue with these approaches is the fidelity of the extracted rules; that is, how faithful the rule-set behaviour is to that of the net.

2.2 Explaining Ensembles

For the black-box approaches described in the previous section the contents of the black-box can be an *ensemble* of neural networks, as easily as a single neural net.

Domingos [8] describes a decision tree-based rule extraction technique that uses the ensemble as an oracle to provide a body of artificial examples to feed a comprehensible learner. Craven and Shavlik [5] describe another decision tree-based rule extraction technique that uses a neural network as an oracle to provide a body of artificial examples to feed a comprehensible learner. Clearly, this technique would also work for an ensemble of neural networks.

The big issue with such an approach is the *fidelity* of the extracted rules; that is, how closely they model the outputs of the ensemble. Craven and Shavlik report fidelity of 91% on an elevator control problem. Emphasising the importance of the ensemble, Domingos reports that his technique preserves 60% of the improvements of the ensemble over single models. He reports that there is a trade-off between fidelity and complexity in the comprehensible models generated; models with high fidelity tend to be quite complex. It is not surprising that comprehensible models that are very faithful to the ensemble will be very complex; and thus less comprehensible.

2.3 Global versus Local Explanation

The focus of this paper is on explaining predictions on a case by case basis. This is different to the current thrust of neural network explanation research.

One author who has also taken this approach is Sima [12] and his approach is reviewed by Cloete and Zurada [4]. Local explanations of time-series predictions have also been explored by Das et al. [7].

Most other researchers have focused on producing global model explanations. These models aim to fully describe all situations in which a particular event will occur. Although this may be useful in many situations, it is argued here that it is not always appropriate. For example, it may be useful in the problem of predicting success in IVF(in-vitro fertilisation) research for instance, studied by Cunningham et al. [6] ,to produce a global model of the phenomenon. Such a model would allow practitioners to spend time understanding the conditions leading to success and to focus their research on improving their techniques. Also, a global model would allow the targeting of potential recipients of the treatment who have a high probability of success. This would lead to a monetary saving for the health service and would avoid great disappointment for couples for whom the treatment would most likely fail. A global model might also allow doctors to suggest changes a couple might make in order to improve their chances of success with the treatment.

In the accident and emergency department of a busy hospital, the explanation requirement would be quite different. Here the need is for decision support rather than knowledge discovery. What is needed is an explanation of a decision in terms of the symptoms presented by individual patients.

3 System Description

3.1 Datasets

Two datasets were used in the analysis presented in this paper. Since the objective of the research is to produce explanations of predictions the main evaluation is done on some Bronchiolitis data for which we have access to a domain expert. This data relates to the problem of determining which children displaying symptoms of Bronchiolitis should be kept in hospital overnight for observation. This data set comprising 132 cases has a total of 22 features, composed of 10 continuous and 12 symbolic and a single binary output reflecting whether the child was kept overnight or not.

In order to provide some insight into the operation of the system we also include some examples of explanations for the Iris data-set [2]. This is included to show graphically the types of rules that are selected by the system.

3.2 Neural Networks

The neural networks used in this system are standard feed-forward networks trained using backpropagation. It is well known that although neural networks can learn complex relationships, they are quite unstable. They are unstable in the sense that small changes in either the structure of the network(i.e. number of hidden units, initial weights etc.) or in the number of training data may lead

to quite different predictions from the network. An effective solution is to use a group (ensemble) of networks trained to approximate the same function, and to aggregate the outputs of the ensemble members to produce a prediction [6,9].

One technique for dividing the data and combining the networks is bagging [3] (short for bootstrap aggregating). This involves randomly selecting examples with replacement from the full set of data available for training. If the size of these bootstrap sets is the same as the full training set, roughly a third of the examples will not be selected at all for each individual sample. These remaining samples can be used as a validation set to avoid overfitting the network to the data. For regression tasks Breiman [3] simply takes the average of the individual network outputs as the ensemble output. For the classifications tasks used in this evaluation, majority voting is used to determine the ensemble prediction.

Ensembles have the added benefit that in reducing the instability of networks the prediction performance is also improved by averaging out any errors that may be introduced by individual networks. The more unstable the networks, the more diverse the networks and thus the greater the improvement of the ensemble over the accuracy of the individual networks.

3.3 Rule Extraction

The approach to explaining ensembles of neural networks that we describe here involves extracting rules from the individual networks in the ensemble, finding the rules that contribute to the prediction and selecting the rules that best fit the example. The approach we use for rule extraction is a fairly standard black-box approach - similar to that used by Domingos [8]. One major difference between our approach and that of Domingos is that Domingos built a single tree based on the results of using the ensemble as an oracle. We also implemented this solution and compared it with our approach; the results of both approaches are included in the evaluation. Our rule extraction process uses the neural networks as *oracles* to train a decision trees using C4.5 [10]. C4.5Rules is then used to extract rules from this decsion tree. The steps are as follows:

1. Generate artificial data by small perturbations on the available training data.
2. Use the neural network to predict an output (i.e. label) for this data.
3. Use this labeled data to train a C4.5 decision tree.
4. Extract rules from this decision tree using C4.5Rules

This yields a set of rules that model the neural net with reasonable fidelity. This number of rules actually produced can be controlled by setting the pruning parameter in the process of building the tree.

3.4 Rule Selection

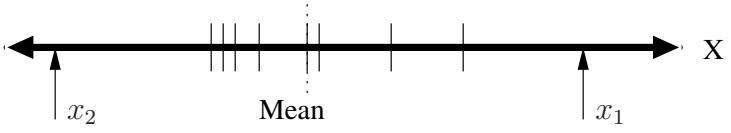
After training an ensemble of networks and building decision trees to model the behaviour of the individual networks we are left with a group of rule-sets, one for each network. The task then is to find the most predictive of these rules for a given input. This is accomplished by executing the following steps:

- Apply each of the rule-sets to the example to produce a prediction from that rule-set
- The rule-sets vote among themselves to decide the overall ensemble prediction
  - Any rule-set that did not vote for this predicted outcome is now discarded
  - Rules that did not vote for the winning prediction within the remaining rule-sets are also discarded
  - This leaves only rules that contributed to the winning prediction

It is from this subset of relevant rules that the most relevant rules will be chosen. In order to select the most relevant rules, it is first necessary to know some statistics about these rules. These are computed after initially producing each rule set.

Rule Coverage Statistics After producing each rule-set, it is necessary to propagate each data item in the set of data used to train the network through each rule. If a rule fires for a particular example and both the example and the rule have the same target, then this example is saved with the rule. The number of examples saved with the rule is considered to be the coverage for that rule.

However, it is possible to go beyond a simple coverage figure. This is done by analysing the individual rule antecedents with respect to the examples covered. For each antecedent in the rule that tests a numeric feature, the mean and standard deviation of the values of that feature contained within the examples covered by that rule can be calculated. The is shown graphically for a single feature in Figure 1. For antecedents testing symbolic features, a perfect fitness score is automatically assigned since any example firing that rule must by definition have the value of that symbolic feature.



Rule antecedent: $x_2 < \mathrm{X} < x_1$

Fig. 1. Number line showing limit of rule antecedent test and several values from examples that fired this rule

Having calculated the above statistics for each of the antecedents, it is now possible to calculate the "fit" of future unseen examples to each of the rules.

Firstly a fit is calculated for each of the numeric features in the rule. This is calculated using equation 1.

$$\mathrm{Fit}_{\mathrm{X}} = \left| \frac{x - \mu}{\sigma} \right| \tag{1}$$

The antecedent with the maximum (i.e. poorest) fitness score is then selected as the fit for the rule as a whole. This is similar to the approach taken in MYCIN [11] as shown in equation 2.

$$\mathrm{MB}[h_1 \wedge h_2, e] = min(\mathrm{MB}[h_1, e], \mathrm{MB}[h_2, e]) \tag{2}$$

In this case the measure of belief(MB) in two terms in conjunction in a rule would be the MB of the weaker term.

Finally, basing the fitness on the distance from the mean is not appropriate in situations where a term is only limited on one side (e.g the first example in section 4.1). In those situations, an example with a feature value on the far side of the mean to the limit is given the maximum fitness, i.e. it is considered to fit the rule well.

3.5 Rule Selection Policies

This fitness measure gives us our main criteria for ranking rules and, so far, has proved quite discriminating in examples examined. However in the Bronchiolitis scenario (see section 4.2) ties can occur when a group of rules all have maximum fitness. Ties can be resolved in these situations by considering rule specificity, i.e. the number of terms in the rule. In situations where simple explanations are preferred, rules with few terms are preferred. In situations where elaborate explanations might be interesting rules with more terms in the left-hand-side can be ranked higher.

The doctor examining the results of the Bronchiolitis data suggested that, in practice, simple explanations might be appropriate for holding a patient overnight whereas more elaborate explanations might be necessary for discharge. The logic behind this is that a single symptom might be enough to cause concern about a child whereas to discharge a child no adverse symptoms should be observable.

So in selecting and ranking rules to explain the Bronchiolitis data the main criterion was the ranking based on the rule fit. Then ties were resolved by selecting the most simple rules for admissions and the most complex rules for discharges. This produced very satisfactory results.

4 Evaluation

Evaluation of this research is not straightforward. To appreciate the quality of the suggested rules, it is necessary to have a good understanding of the domain under investigation. For this reason, the results generated for the Bronchiolitis dataset were given to an expert in this area and his opinions are recorded in section 4.2.

For each of the datasets a total of ten examples were held back from training of the networks and used for evaluation only. For each one of these examples the five most predictive rules were chosen. Also included in the results was a second

set of rules selected from rules that were extracted from a decision tree that was trained to model the behaviour of the vote over the ensemble of networks. This second set of results was included as a comparison to see if the system could select more accurate rules given the more diverse rule-sets of the ensemble members.

4.1 Iris Dataset

In order to offer some insight into the operation of the system, we can show some examples of it in operation on the Iris dataset [2]. The Iris data contains three classes and four numeric features so the rules are much simpler than those produced for Bronchiolitis. This dataset is so simple in fact that the fidelity results are close to perfect. In order to make the problem somewhat more difficult(and to produce more diverse ensemble members) the total number of examples for each class was cut from 50 to 20. A plot of the training data in two dimensions is shown in Figure 2.

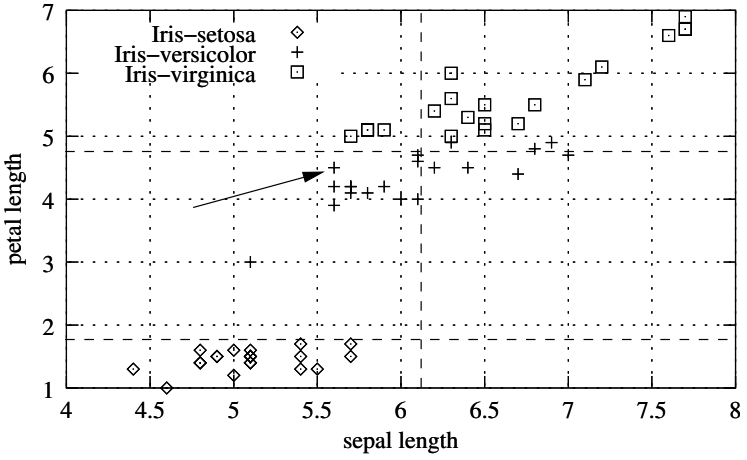


Fig. 2. Iris data plotted in two dimensions

From Figure 2 we can see how the two following rules were selected to explain two different examples. The number in the square brackets preceding the rules is the fit for the example for that rule. The first rule classifies an Iris-setosa. The zero fit indicates that the example being tested fell on the far side of the mean for the rule and hence was given a maximum fit as described in section 3.4. The second rule classifies the example indicated by the arrow in the figure. It has a fit of 0.76 because this example is actually quite close to the limit for the second term for that rule. The example fits the first term well but the poorer fitness is

chosen as the overall fitness for the rule. Nevertheless this rule was selected as the best rule from and ensemble of 9 members.

```
[0.000000]
IF petal_length <= 1.874200
 THEN Iris-setosa

[0.759346]
IF sepal_length <= 6.120790
 AND 1.874200 < petal_length <= 4.797420
 THEN Iris-versicolor
```

4.2 Bronchiolitis Dataset

In the case of the Bronchiolitis dataset Paul Walsh, a doctor in Crumlin Children's Hospital Dublin and the original provider of the data wrote a response to each of the selected rules for each of the tested examples.

Before analysing some of the comments of the expert, some statistics are included in Table 1 describing important characteristics of the network and rules. These statistics were calculated using 10 fold cross validation with an ensemble of five neural networks per fold. The accuracies for each network and its associated extracted rules were calculated on the remaining data in each fold. The fidelity between the network and rule results was also calculated. Finally the results from all the individual networks in a single fold were combined using voting to produce an accuracy for the ensemble.

Table 1. Average and standard deviation figures for the accuracy and fidelity using 10-fold cross validation on Bronchiolitis data

Average Ensemble Accuracy	$73\% \pm 9$
Average Network Accuracy	$69\% \pm 12$
Average Rule-set Accuracy	$67\% \pm 12$
Average Network/Rules Fidelity	$84\% \pm 9$

The fidelity result in Table 1 is of particular interest. The fidelity figure is a measure of how well the rules actually model the network behaviour. This measure is estimated by executing both the network and the rules with all the data. The fidelity is the percentage of times the rule-sets agree with the associated network. Clearly it is important that this figure be as high as possible.

The results in Table 1 show that we get a reasonable improvement in accuracy from using ensembles in the Bronchiolitis dataset. We also get increased stability, the individual network results in the ensembles varied more than the ensemble results.

In general many of the rules considered 'excellent' in the sets presented to the expert were among the first presented(i.e. those with the greatest fitness). From this it would appear that the fitness criterion was effective in selecting good rules. An example of one of these rules is included below:

```
IF Dehydration = None
 AND Retractions = None
 AND 92.397100 < SaO_2_2
 AND BS <= 0.358798
 THEN DISCHARGE
```

In the rule above there are two tests on numeric features and two on symbolic features. The fitness is influenced by the numeric features since the fitness on symbolic features will be 'perfect' if the rule applies to the example.

In more detail, the domain expert was asked to examine and rate rules explaining 10 examples. At most five ranked rules were presented to the expert as good explanations for the prediction(some examples had fewer than five rules that fired from all rule-sets in the ensemble). In addition to these ranked rules, rules that comprised solely of antecedents testing symbolic features were also presented. Rules comprising symbolic features only will automatically have perfect fitness. In total there 60 rules were presented to the expert to explain the 10 examples and 90% of these were correct explanations.

A very small number of rules were marked as definitely being incorrect(4 examples had wrong rules). Of the remaining rules, all contributed in aiding the explanation of the prediction according to the expert. In just eight rules, one of the antecedents in the rule selected did not add much to the knowledge contained within the rule, although the rule as a whole was still considered acceptable. Of the remaining rules, six were marked as excellent, indicating that those rules described almost exactly the published criteria for decision making and covered all of the most important features in a single rule.

By comparison, the rules selected from the set of rules derived from the decision tree built to model the ensemble as a whole were less useful. There were only seventeen rules in total(there were far fewer rules in the single ruleset to select from) and of these only two were marked as excellent and three contained wrong or misleading antecedents.

It is interesting to note that in both cases above, the rules with antecedents comprising tests of only symbolic features are only once marked as excellent and once marked as wrong. Most of the rest of these rules were described as "common sense" by the expert.

A final note of interest is that for predictions to send a child home, the explanations given were consistently more accurate(this was true for both the rules selected from the individual networks and from the rules derived from the ensemble targets). This could be due to the fact that any child that goes home, is more likely to display very well defined symptoms. While a child whose symptoms may not have reached critical levels is admitted because the doctor knows through intution that the child will soon display those levels.

4.3 General Observations on Medical Datasets

Some final points should be noted about medical datasets in general that could lead to skewed results:

- Subjective features may exist where the opinions of those collecting the data may have differed.
- Several of the examples in the dataset may have been influenced by environmental factors that cannot be expressed in the data and may have been responsible for a prediction that otherwise wouldn't normally be the case. For example, in the Bronchiolitis scenario there might be concern about the home environment into which a child might be discharged.

Both of these facts are, for all practical purposes, unavoidable in medical datasets and present a particular challenge to the researcher during their analysis.

5 Conclusions & Future Work

These results encourage us that explanations built from rules derived from component neural networks will be more insightful that rules derived from the ensemble as a whole. This work was based on the hypothesis that the effectiveness of ensembles depends on members of the ensemble specialising in different regions of the problem space. Thus, an explanation of a prediction of an ensemble for an individual example needs to seek out this specialising member. Explanations based on viewing the ensemble as a black-box will be more bland. This preliminary evaluation seems to support this.

The process of rule ranking based on the fitness criterion described here is not yet a complete solution. Problems still exist for rules that have only symbolic features since these will automatically get maximum fitness and this will often not be appropriate. There is also the potential for rules with numeric features to have maximum fitness as explained in section 4; however, the use of antecedent specificity as a further criterion to address this issue shows promise.

It became clear during the evaluation that features that were not strongly predictive were turning up in rules where they were not useful. Because of this we propose to precede the whole process with a feature selection process to weed out poorly predictive features. In general, it seems to us wise to precede any explanation exercise with feature selection since it will relieve the explanation process of the burden of accounting for features that are not very relevant.

References

1. Andrews, R., Diederich, J. & Tickle A. (1995) A Survey and Critique of Techniques For Extracting Rules From Trained Artificial Neural Networks, *Knowledge Based Systems 8*, pp373-389. 451
2. Blake, C. L. & Merz, C. J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science. 452, 456

3. Breiman, L., (1996) Bagging predictors, *Machine Learning*, 24:123-140. 449, 453
4. Cloete, I., & Zurada J. M., (2000) *Knowledge Based Neurocomputing*(MIT Press, Cambridge, Massachusettes). 452
5. Craven, M., & Shavlik, J., (1999) Rule Extraction: Where Do We Go from Here?, University of Wisconsin Machine Learning Research Group Working Paper, 99-1. 451
6. Cunningham, P., Carney, J., Jacob, S., (2000) Stability Problems with Artificial Neural Networks and the Ensemble Solution, *AI in Medicine*, pp217-225, Vol. 20, No. 3. 449, 452, 453
7. Das G., Lin K., Mannila H., Renganathan G., Smyth P., (1998) Rule Discovery from Time Series, *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, (AAAI Press). 452
8. Domingos P., (1998) Knowledge Discovery via Multiple Models, *Intelligent Data Analysis*, 2, 187-202. 451, 453
9. Hanson, L. K., Salomon, P., (1990) Neural Network Ensembles, *IEEE Pattern Analysis and Machine Intelligence*, 1990. 12, 10, 993-1001. 453
10. Quinlan, J. Ross., (1988) *C4.5 Programs for Machine Learning*(Morgan Kaufmann Publishers Inc., San Mateo, CA). 453
11. Shortliffe, E. H., (1976) Computer-Based Medical Consultations: MYCIN, New York, Elsevier. 455
12. Sima, J., (1995) Neural Expert Systems, *Neural Networks* 8(2) pp261-271. 452
13. Wall, R., Cunningham, P., (2000) Exploring the Potential for Rule Extraction from Ensembles of Neural Networks, 11th Irish Conference on Artificial Intelligence & Cognitive Science (AICS 2000), J. Griffith & C. O'Riordan (eds.) pp52-68 (also available as Trinity College Dublin Computer Science Technical Report TCD-CS-2000-24). 451
14. Zenobi G., & Cunningham P., (2001), Using Diversity in Preparing Ensembles of Classifiers Based on Different Feature Subsets to Minimize Generalisation Error, *12th European Conference in Machine Learning (ECML 2001)*, eds L. De Raedt & P. Flach, LNAI 2167, pp576-587, Springer Verlag. 450