Monograph:
Zhe , Ma. and Harrison, R.F. (1995) An "Artificial Expert"-Knowledge Acquisition via Neural Networks. Research Report. ACSE Research Report 578 . Department of Automatic Control and Systems Engineering

---

An "Artificial Expert"
— Knowledge Acquisition via Neural Networks

Zhe Ma, Robert F Harrison
The University of Sheffield,
Department of Automatic Control and Systems Engineering,
Mappin Street, Sheffield, S1 3JD, UK
E-mail: z.ma@shef.ac.uk, R.F.Harrison@shef.ac.uk
Telephone: 44-114-2825198 Fax: 44-114-2731729

Research Report No 578

Abstract

Artificial Neural Networks (ANNs) perform adaptive learning. This advantage can be used to solve knowledge acquisition bottle-neck in knowledge engineering by rule extraction from the ANNs. This paper proposes a rule extraction method combining both open-box (white-box) and black-box approaches to analyse a trained Multilayer Perceptron in order to extract general production rules accurately, abstractly and efficiently.

Key Words: Rule Extraction, Hybrid Knowledge-based System, Neural Network, Knowledge Acquisition, Attribute Selection

1. Introduction

Knowledge Acquisition (KA) is the bottle-neck in knowledge engineering. KA is a process to transfer the domain knowledge into an explicit declarative knowledge formula such as a set of rules. The quality of the acquired knowledge depends on many aspects, such as the availability of the domain experts, their expertise, their ability and attitude to express their expertise, the relationship between the knowledge engineers and the domain experts, and so forth. KA is a time-consuming and not generally reliable process in knowledge engineering. A efficient and reliable KA method is a crucial factor to convince users accepting AI technology, such as expert systems.

Artificial Neural Networks perform adaptive learning on the given pattern set, instead of being told the domain knowledge. This shifts the burdens to one of high quality data gathering. If the knowledge acquired by the ANNs can be translated into symbolic forms, such as production rules, it may then be used to rule-based expert systems. This approach is called *Artificial Expert* because an ANN plays the role of the group of domain experts and knowledge engineers in the KA process.

Rule extraction methods from ANNs can be categorised depending on whether or not they correspond to the black-box or to the open-box (white-box) approach. In the process of rule construction, the open-box approach is concerned with the internal connections of an ANN, and the black-box approach only takes account of the input/output correspondence of the ANN. Domain knowledge has been encoded in the ANN during training, in a subsymbolic form. Analysis of this knowledge in a quantitative way, which is required by knowledge-based systems, is extremely difficult owing to the parallel, distributed properties, the effect of the control parameters, and the mutual dependence of the substructures of the ANN. This is the problem that rule extraction technology using the open-box approach are facing. In order to simplify this problem, some methods alter the trained ANN by pruning or by grouping the connections, adding more layers of units, or retraining the ANN[2][6]. These are usually very time-consuming and the generality of the ANN is not usually guaranteed in such cases. Methods using the black-box approach, on the other hand, must face the problem of combinatorial explosion, as does [5]. To overcome this, some arbitrary protocol is usually adopted to restrict the size of the search space, but this in turn may have an adverse effect on the extracted rules.

The rule extraction method described here combines both the black-box and the open-box approaches, aiming to collect maximal information with as small a computational cost as possible. In the open-box approach, it turns out to be simple to obtain some qualitative properties of the domain knowledge, such as the statistical relationship between the input and the output variables. In the black-box approach, we try to correlate the inputs and outputs of the ANN restrictively in the context of the training set, rather than of the whole possible space. Although neither the qualitative knowledge from the open-box approach nor the quantitative knowledge from the

black-box approach alone is usually sufficient, they complement each other for improving the quality of the rule extraction process. We have generated a hybrid knowledge-based system, GR2, which integrates an ANN and a rule-based system via the rule extraction interface [4].

We take a Multilayer Perceptron (MLP) as the ANN example in this paper. The paper is organised as follows. Sections 2 and 3 define the attribute selection criteria in the open-box and the black-box approaches respectively. We use the term "attribute selection" since the rule extraction is a Machine Learning technology. Section 4 describes the rule extraction algorithm. Full details of the method can be found in [3] and [4]. The method has been used on a real-world medical data set of breast cancer diagnosis which is described in Section 5. A summary is given in the last section.

2. Attribute Selection Criterion 1: Potential Default Set (PDS)*

For the open-box approach, the *Potential Default Set* specifies a set of input variables which are potentially unimportant in deriving the network's output. This is obtained in two steps.

(1) Calculate the *Static Link* between each input unit and each output unit of the MLP $L_{oi} = \sum_h w_{ih} \cdot w_{ho}$, where w_{ih} is a connection from an input unit I_i to a hidden unit H_h in the MLP, w_{ho} is one from the hidden unit H_h to an output unit O_o. The static link represents the statistical quality of the I/O correspondence regardless of the presence of any unit's activation.

(2) In each pattern, select those input variables satisfying the conditions in Table 1

Table 1: Conditions when I_i may be ignorable

L_{oi}	O_o	I_i
≥ 0	$[1-\delta, 1]$	0
≥ 0	$[0, \delta]$	1
≤ 0	$[1-\delta, 1]$	1
≤ 0	$[0, \delta]$	0

where δ is the error tolerance for the classification by the MLP [4]. These conditions are given under the assumption that if the static links always reflect the correspondence between the input and output units involved, switching those input variable values will not change the output status. The strengths of the static links are not important here. However, the assumption does not always hold for nonlinear problems, so a further stage is required.

3. Attribute Selection Criterion 2: Feature Salient Degree (FSD)

In the black-box approach, the *Feature Salient Degree* is defined to represent the dependence degree of the output variable on an input variable in each pattern. The FSD is a $P \times N$ matrix, where N is the number of input variables and P is the number of the patterns in the training set. First, we define the fsd† matrix, whose jith element is

$$fsd_{ji} = \sum_{\{k \mid (j \neq k,\ o_o^j \neq o_o^k,\ I_{ji} \neq I_{ki})\}} \frac{1}{|P_j, P_k| \cdot 2^{|P_j, P_k|}} \qquad (1)$$

and the FSD matrix is defined thus:

$$FSD = \frac{fsd}{\max(fsd)} \qquad (2)$$

In equation (1), P_j and P_k denote different patterns in the training set. $|P_j,P_k|$ is the Hamming distance of the input vectors of the patterns. (1) takes account of those patterns, P_k, in the training set whose input variable I_i and the output variable are different from those in P_j. If we define $I(P_j)$ as the input vector of P_j. To each pair of patterns P_j and P_k, the subset $I(P_j)$-$I(P_k)$ is regarded as the subset determining the class which P_j's output represents. The population of those possible input vectors subsuming this property set in the input space $\{0,1\}^N$, is

$$\frac{2^{N-|P_j, P_k|}}{2^N} = \frac{1}{2^{|P_j, P_k|}}$$

Furthermore, assuming each input in the property set equally characterises the property, the contribution of each input variable is $\frac{1}{|P_j, P_k| \cdot 2^{|P_j, P_k|}}$. This explains how equation (1) is formed.

*. We consider rules with only one output variable. If a domain possesses more than one output variable, it can be partitioned into subproblems, each of which has single output variable.

†. This definition slightly differs from those given in our previous papers [3][4]. Although the results using the different definitions for fsd are similar, this one has the best intuitive explanation.

The fsd matrix is normalised by its maximal element in equation (2) so that all its elements lie in [0,1], since fsd is a non-negative matrix. This operation regularises the matrix when the training set is redundant or incomplete to the whole input space.

4. Rule Extraction Algorithm

The kernel rule extraction algorithm uses the two attribute selection criteria on each training pattern P_j

(1) Set a control parameter $\tau \in (0,1]$.

(2) Collect all input variables $\psi=\{I_i \mid FSD_{ji} \geq \tau$ or $(I_i \notin$ PDS and $FSD_{ji} \geq \tau/(N \times \log N)\}$

(3) Each minimal subset $\theta_k \subset \psi$ constructs the premise part of a new rule, and the output variable forms the consequence of the rule. The "minimal subset" means θ_k is not subsumed by other subsets.

In step 2, $FSD_{ji} \geq \tau/(N \times \log N)$ is used to filter out those I_is having very small FSD values. This greatly reduces the computation to find subsets in step 3. In practice, we can define separate control parameters for classes respectively, as shown in the next section.

The algorithm also includes other operations such as rule pruning and generalisation. The extracted rule set will then be processed by a rule validation and a rule verification processes, resulting in a representation of the domain knowledge under uncertainty. These details have been described in [4].

5. Experiment on Breast Cancer Diagnosis Records

The rule extraction method has been successfully applied to some traditional artificial problems such as the two or more bit AND, OR, and parity problems, and to several medical domains. We report here the experiment of the method to breast cancer diagnosis. The data set consists of 413 patient records, each comprising ten binary-valued symptoms (inputs) recorded from human-observation of breast tissue samples, together with the actual outcome for each case, which is a lesion proved to be either malignant or benign. The data set possesses a certain level of noise: there are only 188 distinctive cases including 12 conflicting cases. Further details can be found in [1].

There are three performance useful indicators used in the domain. *Sensitivity* is defined as the ratio of the number of correct positive diagnoses to the number of positive outcomes. This is important as the disease is life-threatening. *Specificity* is defined as the ratio of the number of correct negative diagnoses to the number of negative outcomes. This is important as treatment is expensive and can be risky. *Accuracy* is defined as the ratio of the number of correct diagnoses to the total number. In this domain, specificity of benign must be high, to avoid unnecessary surgery being carried out.

The data were divided into three randomly selected sets: 100 cases as the training set, the next 100 cases as the verification set, and the remaining 213 cases as the test set. A conventional MLP with one hidden unit layer was trained on the training set. And then test it on the verification set in order to find the optimal MLP structure, which was found to be one with 10 input units and 5 hidden units. The results on the verification set are sensitivity=95.4%, specificity=93.5% and accuracy=95%. This MLP was then applied on the test set, whose results are shown in Table 2.

Using the rule extraction algorithm on the optimal MLP and the training set, we chose two thresholds for the two classes, the positive threshold (pt) for the malignant class, and the negative threshold (nt) for the benign class. Both thresholds were varied in the range [0.01, 0.5], increasing in 30 steps respectively. When the rule set was extracted at each of the 30×30 choices of the thresholds, it was applied to the test set. The results of the 900 tests are plotted in Figure 1. We find the optimal choice of the thresholds to be: pt in [0.195,0.212] and nt in [0.144,0.163]. The 900 tests took only 22 minunes 16 seconds on a Sun Sparcstation.

Table 2 compares the results of the extracted rules on the test set with those from the MLP.

Table 2: Results on the Test Set of the Breast Cancer Diagnosis Records

(%)	MLP, with 10:5:1 structure	Rule Extraction, pt=0.2, nt=0.16
Sensitivity	92.9	94.7
Specificity	95	93
Accuracy	93.9	93.9

The extracted rules at the choice pt=0.2, nt=0.16 are given as the follows, each of which was attached a confidence degree, which may be used to assess the level of belief the user might have in any particular rule [4].

IF (~Naked, ~Foamy) THEN (Benign); (0.173913)

IF (Naked, Apocrine) THEN (Benign); (1)

IF (~Foamy, Apocrine) THEN (Benign); (1)

IF (~Necrotic) THEN (Benign); (0.2)

IF (~Foamy) THEN (Benign); (0.2)

IF (Necrotic) THEN (Malignant);	(0.733333)
IF (Size) THEN (Malignant);	(1)
IF (ICL) THEN (Malignant);	(0.777778)
IF (3D) THEN (Malignant);	(0.703704)

The ten binary symptoms (inputs/premises) are defined as:

DYS: True if majority of epithelial cells are dyhesive, false if majority of epithelial cells are in cohesive groups.

ICL: True if intracytoplasmic lumina are present, false if absent.

3D: True if some clusters of epithelial cells are not flat (more than two nuclei thick) and this is not due to artefactual folding, false if all clusters of epithelial cells are flat.

NAKED: True if bipolar "naked" nuclei in background, false if absent.

FOAMY: True if "foamy" macrophages present in background, false if absent.

NUCLEOLI: True if more than three easily visible nucleoli in some epithelial cells, false if three or fewer easily visible nucleoli in epithelial cells.

PLEOMORPH: True if some epithelial cell nuclei with diameters twice that of other epithelial cell nuclei, false if no epithelial cell nuclei twice the diameter of other epithelial cell nuclei.

SIZE: True if some epithelial cells with nuclear diameters at least twice that of lymphocyte nuclei, false if all epithelial cell nuclei with nuclear diameters less than twice that of lymphocyte nuclei.

NECROTIC: True if necrotic epithelial cells present, false if absent.

APOCRINE: True if apocrine change present in all epithelial cells, false if not present in all epithelial cells.

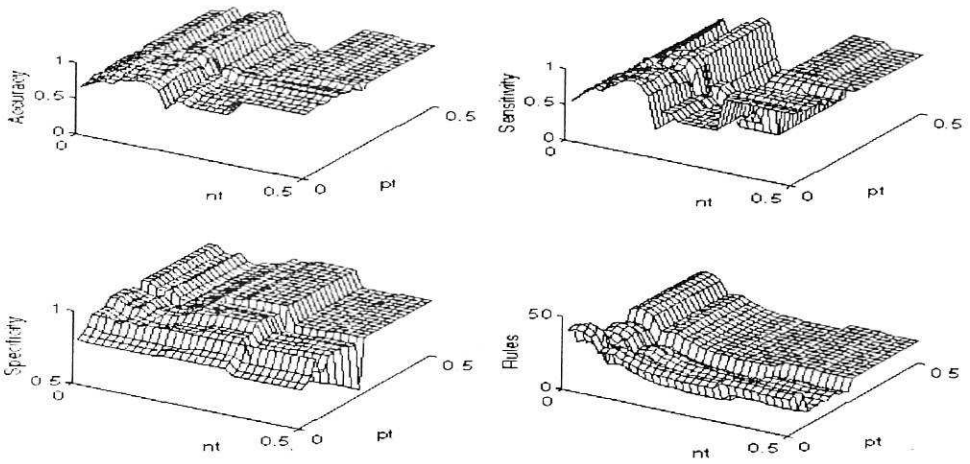

Figure 1. Performance of the Extracted Rules on the test set of Breast Cancer Records as the positive threshold pt and the negative threshold nt are changed in [0.01, 0.5]

6. Conclusion

An optimal method to extract knowledge in symbolic form from ANNs can relieve the KA bottle-neck problem and support knowledge engineering automatisation to a great extent. A combination of the black-box and open-box approaches to rule extraction makes the process efficient, effective and easy to control, resulting in the extracted rules representing the domain knowledge as accurately as the ANNs can learn.

References

[1] J. Downs, R. F. Harrison, S. S. Cross. *A Neural Network Decision-Support Tool for the Diagnosis of Breast Cancer.* Hybrid Problems, Hybrid solutions, IOS Press, Edited by J. Hallam, 1995, pp51-60

[2] L. M. Fu. *Rule Generation from Neural Networks.* IEEE Transactions on Systems, Man, and Cybernetics, 24(8), 1994, pp1114-1124

[3] Z. Ma, R. F. Harrison, R. L. Kennedy. *A Heuristic for General Rule Extraction from a Multilayer Perceptron.* Hybrid Problems, Hybrid solutions, IOS Press, Edited by J. Hallam, 1995, pp133-144

[4] Z. Ma, R. F. Harrison, R. L. Kennedy. *GR2 - A Hybrid Knowledge-based System Using General Rules.* Proceedings of IJCAI-95, 1995, Montreal. To appear

[5] K. Saito, R. Nakano. *Medical Diagnostic Expert System Based on PDP Model.* International Conference on Neural Networks, IEEE press, San Diego CA, 1988, pp255-262

[6] G. Towell, J. W. Shavlik. *Extracting of Refined Rules from Knowledge-Based Neural Networks.* Machine Learning, 1993, Vol 13, No1, pp 71-101