Analysis of Rules Discovered by the Data Mining Process

Ken McGarry
School of Computing and Technology
University of Sunderland, St Peters Campus
St Peters Way, Sunderland, SR6 ODD, United Kingdom
e-mail: ken.mcgarry@sunderland.ac.uk

Abstract
This paper describes how symbolic rules may be extracted from Radial Basis Function neural networks and shows how they can be used by the data mining and knowledge discovery process. Rule extraction overcomes a major disadvantage of neural networks which is concerned with making the comprehensibilty of the learned internal model more open to scrutiny. Having extracted the symbolic rules we show how they are assessed and ranked for interesting or novel features. Two such techniques are presented here, the first is a data driven approach that uses objective mathematical measures to identify interesting patterns or features. The second is a goal driven method that uses subjective measures obtained from the user. The measures are applied to rules extracted from RBF neural networks trained on several data sets including benchmark sets from the UCI repository and a large real-world industrial data set.

Keywords: Rule extraction, neural networks, data mining, interestingness measures.

1 Introduction

Data mining and knowledge discovery from data bases has received much attention in recent years. However, the majority of this work has concentrated on producing accurate models without considering the potential gains to be had from understanding the details of these models. Recent work has to some extent addressed this problem but much work has still to be done [4, 7]. In this paper we outline our own methods of assessing the novelty,usefulness and comprehensibility of rules extracted from the data mining process. In fact one of the most insightful definitions of data mining states that to be truly successful data mining should be *"the process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from databases"* that is used to make crucial business decisions, [2]. Currently, there are two main techniques of assessing the interestingness of discovered patterns. The first uses objective mathematical measures to assess the degree of interestingness, many such measures exist. The second method is to incorporate the users subjective knowledge into the assessment strategy. Each of these approaches has various characteristics and we consider how they may be combined to produce a more robust system [5].

The particular system under investigation is comprised of symbolic rules extracted from Radial Basis Function Neural Networks. Neural networks are are accurate classifiers for low level pattern recognition tasks. Unfortunately, they are effectively "black boxes" because their internal representation is difficult if not impossible to decode by humans. Previous work performed by the author has seen this problem partly solved by extracting $< IF..THEN >$ type rules [8, 9]. However, what this work has been lacking is the automatic analysis of these rules using "interestingness" measures to asssess their novelty,usefulness and comprehensibility.

The remainder of this paper is structured as follows. Section two descibes the architecture of the radial basis function network and how symbolic rules may be extracted from its internal parameters. Section three introduces data mining and knowledge discovery and reviews the interestingness measures used for ranking the rules/patterns identified by the rule extraction algorithm. Section four shows the experimental results from several data sets. Section five dicusses the conclusions and areas for further work.

2 Rule Extraction from RBF Neural Networks

In this section we briefly discuss the architecture and training of Radial Basis Function (RBF) neural networks and the motivation for extracting symbolic rules. Rule extraction is a seen as a solution to the "black box" problem of neural networks whereby their internal structure is difficult to interpret.

2.1 Radial Basis Function Networks

The RBF network consists of feedforward architecture with an input layer, a hidden layer of RBF "pattern" units and an output layer of linear units. The input layer simply transfers the input vector to the hidden units, which form a localized response to the input pattern. The activation levels of the output units provide an indication of the nearness of the input vector to the classes. Learning is normally undertaken as a two-stage process. The objective of hidden layer learning is to locate the radial basis function centres and to determine the optimum field widths in relation to the training samples. The right half of figure 1 shows the RBF architecture. Several schemes for locating hidden unit parameters have been suggested. Broomhead and Lowe used a uniformly distributed lattice of hidden units but this proved to be impractical for high input dimensionality as the number of hidden units grew exponentially [1]. The simplest procedure is to randomly set them as *prototypes* of a subset of the training data. The left half of figure 1 highlights the local nature of each hidden unit as it maps into only a limited part of the input space, thus enabling a single rule to be formed from each hidden unit.

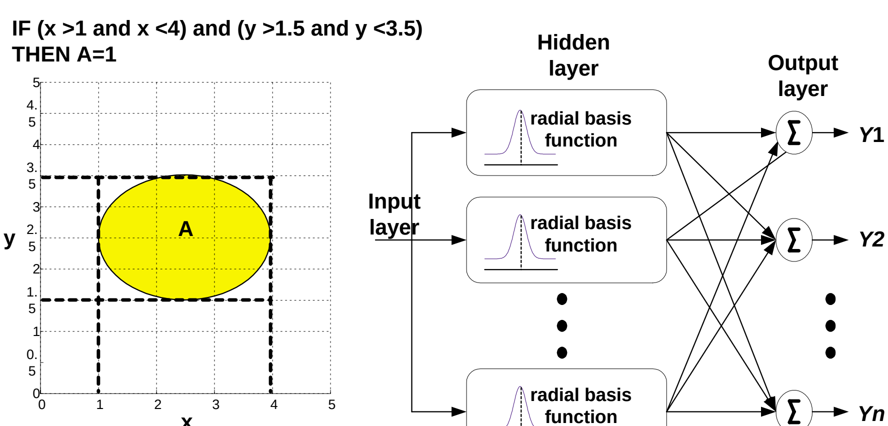

Figure 1: Symbolic approximation of RBF centres

2.2 LREX: Local Rule Extraction Algorithm

The development of the LREX algorithm was motivated by the local architecture of RBF networks which suggested that rules with unique characteristics could be extracted. The LREX algorithm extracts IF..THEN type rules based on the premise that a hidden unit can be uniquely assigned to a specific output class. Therefore, by using the centre locations of the hidden units an input vector could be directly mapped to an output class. Experimental work performed on the simpler data sets tended to reinforce this belief. However, hidden unit sharing occurs within networks trained on non-linear or complex data. This phenomena reduces rule accuracy as several hidden units may be shared amongst several classes [8]. The functionality of LREX algorithm is shown in figure 2.

Input:

Hidden weights μ (centre positions)
Gaussian radius spread σ
Output weights $W2$
Statistical measure S
Training patterns

Output:

One rule per hidden unit

Procedure:

Train RBF network on data set
Collate training pattern *"hits"* for each hidden unit
For each hidden unit
  Use $W2$ correlation to determine Class label
  Use *"hits"* to determine S
  Select S format $\{min, max, std, mean, med\}$
  For each μ_i
    $X_lower = \mu_i - \sigma_i * S$
    $X_upper = \mu_i + \sigma_i * S$
Build rule by:
  antecedent = $[X_lower; X_upper]$
  Join antecedents with AND
  Add Class label
Write rule to file

Figure 2: LREX rule-extraction algorithm

Two rules from the Iris domain are presented in figure 3. The antecedent is formed by calculating the lower and upper bounds for each hidden units 'spread'.

Rule 1 :
IF (SepalLength $\geq$ 4.1674 AND $\leq$ 5.8326) AND
IF (SepalWidth $\geq$ 2.6674 AND $\leq$ 4.3326) AND
IF (PetalLength $\geq$ 0.46745 AND $\leq$ 2.1326) AND
IF (PetalWidth $\geq$ 0.53255 AND $\leq$ 1.1326)
THEN..Setosa

Rule 2 :
IF (SepalLength $\geq$ 5.2674 AND $\leq$ 6.9326) AND
IF (SepalWidth $\geq$ 1.9674 AND $\leq$ 3.6326) AND
IF (PetalLength $\geq$ 3.1674 AND $\leq$ 4.8326) AND
IF (PetalWidth $\geq$ 0.46745 AND $\leq$ 2.1326)
THEN..Versicolor

Figure 3: Extracted rules from Iris domain

The use of information theory was made to reduce the number of antecedents within a rule. This involved implementing the ChiMerge algorithm and discretizing the data sets to simplify the process of calculating the information measure. The information measure or information gain about an input feature represents the importance of that feature to resolve class identity. Features with very low values may be excluded from the extracted rules and thus simplied the rule and aids comprehensibility.

3 Interestingness Measures for Knowledge Discovery

The term 'data mining and knowledge discovery' came into use around the late 1980's and has been an intense area of research activity since then. The reason for the surge in interest was due the vast quantities of data generated by modern companies as part of their daily activity. The means to analyse this data far outstretched the conventional statistical techniques. Clearly an automated approached was needed and techniques from machine learning have been applied to understand, describe and predict future patterns. The analysis and ranking of the discovered patterns may present the greatest challenge of all.

3.1 Objective Measures

The majority of the data mining work concentrates on the discovery of accurate and comprehensible patterns. Whilst this approach will provide the user with a degree of confidence regarding the discoveries it falls far short of the notion of "knowledge discovery". The essential task for data mining algorithms is to search for patterns that are "suprising" in addition to being accurate and comprehensible. Criteria such as rule coverage, rule complexity, rule confidence and rule completeness are becoming used as a measure of the interestingness of the discovered patterns [4].

- Small disjuncts. Inductive systems are generally designed, where possible to produce models with large disjuncts i.e. a rules that will cover as many instances of the training set as possible. This aids the understandibility of the model through producing fewer rules. Unfortunately, for some data sets small disjuncts can make up perhaps 20% of the total number of rules in a classifier [6]. The presence of small disjuncts can indicate a noisy data set and/or some interesting exceptions.

- Class imbalance. Should one or two classes predominate the original data set in terms of numbers then the rules predicting the smaller, less well represented class may represent some interesting exception.

- Complexity. The complexity of a rule was assessed by the number of average number of antecedents within the rule body.

- Size/accuracy. The ratio of the size of the rule set against the accuracy provides an indication of .

3.2 Subjective Measures

Subjective measures are more difficult to devise as these require a domain expert to formulate rules to detect patterns. Discoveries by their very nature are often unexpected and therefore suprising to the users and hopefully actionable [10, 7]. Two measures were used to rank the discovered rules.

- Unexpected consequent rankings. Any rules with different consequents but with highly similar antecedents can indicate areas of potential interest e.g. class differences are not so easily differentiated and may lead to inaccurate classifiers and/or highlight mislabelled examples.

- Unexpected attribute rankings. Any rule containing tests on attributes that are not generally perceived to play a major role in class identification should be brought to the users attention.

4 Experimental Results

The rule sets extracted from the RBF networks were analysed by of the measures discussed in the previous section. Most of the data sets will be familiar to the AI community apart from the Vibration data sets which are concerned with fault diagnosis on industrial machinery.

Table 1: Extracted rule set size and accuracy

Data set	**No Rules**	**Attributes**	**Classes**	**Accuracy %**
Monks1	20	6	2	83
Diabetes	65	8	2	76
Credit	50	15	9	93
Vibration 1	30	9	3	73
Vibration 2	100	20	8	94

Table 2: Objective interestingness measures applied to extracted RBF rules

Interestingness Measure	**RBF Rule Sets**				
	Monks	Diabetes	Credit	Vibration 1	Vibration 2
Small Disjuncts	No	No	Yes	Yes	Yes
Class Imbalance	No	No	Yes	No	Yes
Size/accuracy	4.1	1.1	1.8	2.4	0.94
Complexity	6	8	9	9	20

Rating the rules with subjective measures is more difficult. Knowledge about the relevant domains must be available either from an expert or via documentation. Only the vibration domain had access to the necessary expertise therefore, only the results for this rule set are presented.

Table 3: Subjective interestingness measures applied to extracted RBF rules

Interestingness Measure	**Rule Set**	
	Vibration 1	Vibration 2
Unexpected consequences	2	11
Unexpected antecedents	5	3

In terms of new or suprising knowledge, the vibration rule set provided the best results. The rule extraction algorithm in conjunction with the interestingness measures was able to identify two input parameters not generally thought of as being important for identifing machine faults. The RBF training algorithm had used the most discriminating features for detecting the fault classes which was made explicit by the rule extraction algorithm. The rules were then ranked accordingly. The rules using these parameters are valid, accurate and account for 15-20% of the total number of extracted rules.

5 Conclusions

This paper has presented the results of ranking and the analysis of rules extracted from RBF neural networks using both objective and subjective measures. The interestingness of a rule can be assessed by a data driven approach. Unfortunately, objective measures may still allow uninteresting patterns to emerge from the data mining process. As suspected there is no overall optimum measure that can be applied to all data mining patterns and by their nature are highly domain specific [3]. The use of subjective or goal driven measures should provide a means to discard uninteresting patterns. However, in those domains which are imperfectly understood then developing a suitable set of subjective measures may not be possible. Future work will explore the issue of integrating the data and goal driven measures.

References

[1] D. S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, pages 321–355, 1988.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthursamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI-Press, 1996.

[3] A. Freitas. On objective measures of rule suprisingness. In *Principles of Data Mining and Knowledge Discovery: Proceedings of the 2nd European Symposium,Lecture Notes in Artificial Intelligence*, volume 1510, pages 1–9, Nantes, France, 1998.

[4] A. Freitas. On rule interestingness measures. *Knowledge Based Systems*, 12(5-6):309–315, 1999.

[5] R. Hilderman and H. Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'01), Lecture Notes in Computer Science*, pages 247–259, Hong Kong, 2001.

[6] R. C. Holte, L. E. Acker, and B. W. Porter. Concept learning and the problem of small disjuncts. In *International Joint Conference on Artificial Intelligence*, pages 813–818, Detroit, Michigan, USA, 1989.

[7] B. Liu, W. Hsu, L. Mun, and H. Y. Lee. Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering*, 11(6):817–832, 1999.

[8] K. McGarry, S. Wermter, and J. MacIntyre. The extraction and comparison of knowledge from local function networks. *International Journal of Computational Intelligence and Applications*, 1(4):369–382, 2001.

[9] K. McGarry, S. Wermter, and J. MacIntyre. Knowledge extraction from local function networks. In *Seventeenth International Joint Conference on Artificial Intelligence*, volume 2, pages 765–770, Seattle, USA, August 4th-10th 2001.

[10] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996.