Proceedings of the International Conference on
Pattern Recognition, Informatics and Medical Engineering , March 21-23, 2012

Rule Extraction from Neural networks

- a Comparative Study

M.Gethsiyal Augasta
Department of Computer Applications
Sarah Tucker College, Tirunelveli
Tamilnadu, India
augasta@yahoo.com

T. Kathirvalavakumar
Department of Computer Science
V.H.N.S.N College, Virudhunagar
Tamilnadu, India
kathirvalavakumar@yahoo.com

***Abstract* — Though neural networks have achieved highest classification accuracy for many classification problems, the obtained results may not be interpretable as they are often considered as black box. To overcome this drawback researchers have developed many rule extraction algorithms. This paper has discussed on various rule extraction algorithms based on three different rule extraction approaches namely decompositional, pedagogical and eclectic. Also it evaluates the performance of those approaches by comparing different algorithms with these three approaches on three real datasets namely Wisconsin breast cancer, Pima Indian diabetes and Iris plants.**

Keywords-classification; data mining; decompositional; pedagogical; eclectic; neural networks; rule extraction.

I. INTRODUCTION

Huge amount of data are generated on a daily basis in banks, insurance companies, retail stores and on the internet. Neural networks are used to find patterns in the data and to infer knowledge from them in the form of rules. Sometimes, the meaning of rules produced by the neural network method is hard to understand as they are often considered as black box, though they achieve highest classification accuracy for many classification problems with huge datasets. To overcome this problem there have been significant amount of work devoted to the development of rule extraction algorithms that disclose the black box nature of neural networks.

Rule Extraction, is the process of developing natural language-like syntax that describes the behaviour of a neural network [1]. It changes a black box system into a white box system by translating the internal knowledge of a neural network into a set of symbolic rules [2]. Usually, a rule extraction method is based on sample learning by using some classification method to obtain the classification rules. The techniques used for classification mainly include decision trees [3], neural networks [4], and the genetic algorithms [5]. However, by using the decision tree method, it is difficult to mine large datasets. The standard genetic algorithm can possibly have degenerated phenomenon in the evolution process, and it is unable to express relationships of complementing and competing among rules. Neural network has high tolerance of noisy data and it has an ability to classify patterns on which they have not been trained. Also the classification process of a neural network can be described by a set of simple rules. The rule extraction techniques extract the rules either in conjunctive form (if and then) or in subset selection form (M of N). Researchers have proposed many rule extraction algorithms for the trained neural networks [6-9]. Main goals of the rule extraction algorithm to be, (i) generating symbolic representations that are comprehensible by the domain experts, (ii) producing symbolic representations that accurately model the networks from which they were extracted and (iii) requiring neither special training methods nor restrictions on network architecture.

The aim of this paper is to discuss various rule extraction algorithms that have been developed using three different rule extraction approaches namely decompositional, pedagogical and eclectic and to compare its performance of three recent rule extraction algorithms on three real datasets namely Wisconsin breast cancer, Pima Indian diabetes and Iris Plants.

II. OVERVIEW OF RULE EXTRACTION ALGORITHMS

The rule extraction algorithm searches through the structure of the network and/or the contents of a network's training data, and narrow down values across each input looking for the conditions that make up the rules [2]. Many rule extraction algorithms have been designed to reveal the information concealed in the trained neural networks. Most of these algorithms such as NeuroRule [6], RX [7], GLARE [8] and OSRE [9] work well on data with discrete attributes only. If these algorithms are used for extracting rules from real world classification problems with mixed mode attributes, all the continuous attributes must be discretized. The drawback of discretizing continuous attributes is that the accuracy of the rules extracted from the network may decrease [10]. There exist some algorithms that do not require discretization of continuous input data attributes [10, 11]. The techniques of the rule extraction can be grouped into three approaches namely decompositional, pedagogical and eclectic [12]. The decompositional approach extracts the symbolic rules by analyzing the activation and weights of the hidden layers of the neural network. The pedagogical approach extract rules by mapping the input-output relationships as closely as possible to the way the neural networks understand the relationship. The eclectic approach incorporates some of a decompositional approach with some of pedagogical approach.

978-1-4673-1039-0/12/$31.00 ©2012 IEEE

A. Rule Extraction methods by decompositional approach

Rudy Setiono and Huan Liu [13] have proposed a three phase algorithm to understand a neural network via rule extraction. In this algorithm, firstly a weight decay backpropagation network is built so that important connections are reflected by their bigger weights. Secondly the network is pruned, such that insignificant connections are deleted while its predictive accuracy is still maintained. Then finally, the rules are extracted by recursively discretizing the hidden unit activation values. The decompositional technique NeuroLinear [14] is able to extract oblique classification rules from neural networks with one hidden layer. Kim and Lee [15] have proposed an algorithm based on feature extraction and feature combination. It applies to multilayer perceptron networks with two hidden layers. Sang Park [8] has proposed a generalized analytic rule extraction method for feedforward neural networks. This method extracts rules by directly interpreting the strengths of connection weights in a trained network. Krishnan et al., [16] have proposed a search technique for rule extraction from trained neural networks. By sorting the input weights to a neuron and ordering the weights suitably, it finds the combinations of inputs that make the neuron active. FERNN [17] is a decompositional algorithm for classification rule extraction from feedforward neural networks. The extraction process returns oblique rules, but under certain circumstances these rules can be simplified to M-of-N rules or DNF-rules. Tsukimotos method [18] is a decompositional approach that extracts low order rules from each of the node. Odajima et al., [19] have proposed a GRG method for generating classification rules from a dataset with discrete attributes. Neural networks with one hidden layer are trained and the GRG algorithm is applied to their discretized hidden unit activation values. The effectiveness of this proposed method is shown by the application of GRG on three medical datasets with discrete attributes. Setiono et al., [20] have proposed a Recursive Rule Extraction algorithm (Re-RX) to generate classification rules from datasets that have both discrete and continuous attributes. This algorithm is recursive in nature and it generates hierarchical rules. Rule conditions involving the discrete attributes are disjoint for those involving continuous attributes. Ozbakir et al.,[21] have suggested a novel algorithm for classification rule extraction from ANNs. This algorithm employs differential evolution (DE) algorithm for training and touring ant colony optimization algorithm (TACO) for generating classification rules. It extracts rules from a network in a neuron-by-neuron series of steps.

B. Rule Extraction methods by pedagogical approach

Saito and Nakano have proposed a medical diagnosis expert system based on multilayer ANN [22]. It extracts rules by observing the effect on the network output by changing the inputs. Sestilo and Dillon have proposed the BRAINNE system which uses the pedagogical technique [23]. It extracts rules from ANN using backpropagation. A major innovation of the BRAINNE system is, it doesn't require dicretization but it has the capability to deal with continuous data as input. Thrun has proposed the VIA method to extract the rules that map input directly to output [24]. It has used a generate-and-test procedure to extract symbolic rules from the neural network trained by backpropagation, which has not been specifically constructed to facilitate rule extraction. Similar to sensitivity analysis it characterizes the output of the trained ANN by performing the systematic variations in the input pattern. Hayward and Diederich have proposed a method called RULENEG [25]. It extracts rules from trained ANN by stepwise negation. It works with binary inputs and extracts rules in the form of disjunction and conjunctions. The worst case of this algorithm is the number of rules equal to number of training patterns. Craven and Shavlik have proposed a method called TREPAN [26]. It extracts rules in the form of decision tree using sampling and queries. Interval Analysis (IA) method proposed by Filer et al., extracts rules in the form of M-of-N [27]. Saad et al., [28] have proposed a new explanation algorithm HYPINV which relies on network inversion; i.e. calculating the ANN input which produces a desired output. Saad et al., have said that HYPINV may be the only pedagogical rule extraction method, which extracts hyperplane rules from continuous or binary attribute ANN classifiers. Augasta and Kathirvalavakumar [29] have proposed a new rule extraction algorithm called RxREN which extracts classification rules from the trained neural networks using pedagogical approach. The algorithm relies on reverse engineering technique to prune the insignificant input neurons and to discover the technological principles of each significant input neuron of neural network.

C. Rule Extraction methods by eclectic approach

Craven and Shavlik have proposed the Rule-Extraction-as-Learning (REAL) method [30]. It views the rule extraction as a learning task where the target concept is the function computed by the network and the input features are simply the network's input features. Tickle et al., have proposed a general method called a DEDEC [31]. It extracts the set of rules efficiently from a set of individual cases. For rule extraction it identifies the minimal set of information required to distinguish a particular object from other objects. Keedwell et al., [32] have proposed a system to search for rules in the ANN input space using a genetic algorithm. Kaikhah and Doddameti [33] have proposed the method to discover the trends in large datasets using neural networks. It treats the neural network as black box for knowledge discovery but examines the weights for pruning and clustering the hidden unit activation values. It uses the control parameters for data analysis for controlling the probability of occurrences and accuracy of rules. Hruschka and Ebecken [34] have proposed a method Rex-CGA to extract rules from multilayer perceptrons. It is a clustering based approach that employs CGA to find clusters of hidden unit activation values and generates logical rules from these clusters. Kahramanli and Allahverdi [35] have proposed a method to extract rules from trained adaptive neural networks using artificial immune systems. This method is based on both decompositional and pedagogical approaches.

The process of decompositional approach rule extraction is tedious and result in complex and large descriptions. So the drawbacks of this approach are time and computation limitations. But the pedagogical approach algorithms can be faster than the decompositional, since it doesn't analyze the weights or internal architecture of the network, but they are somewhat less likely to accurately capture all of the valid rules describing the network's

behavior. The Eclectic approach algorithms may be slower but more accurate than pedagogical as it combines both decompositional and pedagogical approaches.

III. EVALUATION CRITERIA

The quality of the extracted rules can be measured by several factors such as accuracy, fidelity and comprehensibility. Rule accuracy can be defined as a percentage of data examples that are correctly classified by the extracted rules and it is computed as below:

$$\text{Accuracy} = \frac{\text{Total number of correctly classified examples}}{\text{Total number of examples}}$$

Fidelity defines the ability of the extracted rules to mimic the behaviour of the network from which they are extracted and identifies how much the extracted rules give insight about the ANN architecture. Moreover it is the measure of the agreement between the pruned network and the extracted rule set for correctly classifying the data instances. The extracted rules must not only be accurate, they must also be understandable. The understandability of the rules can be measured by two parameters such as Global comprehensibility and Individual comprehensibility. Large numbers of rules are difficult to understand. So the rule set size is used to measure the global comprehensibility of extracted rule set. Individual comprehensibility will be measured by the number of attributes in each individual rule in rule set. Finally two other criteria are the complexity of the algorithm which is measured by the number of steps needed to extract the rule base, and the complexity of the extracted rule set which depends on the number of rules extracted.

IV. PERFORMANCE COMPARISIONS

This section compares the performance of three recent rule extraction algorithms in study namely GRG [19], RxREN [29] and Rex-CGA [35] using three WEKA's datasets namely Wisconsin breast cancer (wbc), Pima Indian diabetes (pid) and iris plants dataset (iris). The rule extraction algorithms that have been selected for comparisons are in three different rule extraction approaches namely decompositional, pedagogical and eclectic. The detailed descriptions of datasets that are used to compare the performance of the algorithms are listed below and also summarized in Table I.

A. Wisconsin-breast-cancer dataset (breastw):

This dataset was designed to diagnose breast tumors as either benign or malignant. It contains 699patterns and each pattern consists of 9 real value attributes as an input vector and two classes as an output vector. Out of 699 patterns 458 are benign patterns and 241 are malignant patterns.

B. Pima Indians Diabetes dataset (pid):

The problem posed here is to predict whether a patient would test positive or negative for diabetes according to the criteria given by World Health Organization (WHO). This is a two class problem with class value 1 and 2 interpreted as negative and positive results for diabetes. There are 500 patterns of class 1 and 268 of class 2. There are 8 attributes for each pattern.

C. Iris Plants dataset (iris):

Irises are classified into three categories: setosa, versicolor and verginia. Each category has 50 patterns and each pattern possesses four attributes namely sepal length, sepal width, petal length and petal width.

TABLE I. PROPERTIES OF THREE REAL DATASETS

Dataset	**Total examples**	**No. of Classes**	**No. of Attributes**
wbc	699	2	9
pid	768	2	8
irs	150	3	4

Table II shows the comparison results of three algorithms in study on three datasets namely wbc, pid and iris using the results in [19, 29, 35].

TABLE II. RESULT COMPARISON OF THREE ALGORITHMS

Dataset	**GRG**		**RxREN**		**Rex-CGA**	
	Acc (%)	***#rules***	***Acc (%)***	***#rules***	***Acc (%)***	***#rules***
wbc	95.96	2.0	96.4	2.0	96.35	3.0
pid	-	-	77.2	2.0	66.23	5.0
irs	97.3	3.0	97.3	3.0	96	3.0

Regarding the classification accuracy, the pedagogical approach rule extraction algorithm RxREN achieves higher accuracy for all datasets. Also it achieves best rule comprehensibility for all datasets. Though the eclectic approach rule extraction algorithm Rex-CGA achieves higher or similar accuracy for wbc dataset, the rule comprehensibility of the algorithm is lower than other algorithms in the comparison. The rule comprehensibility of decompositional approach rule extraction algorithm GRG is similar to RxREN, but RxREN achieves higher classification accuracy on wbc dataset. Fig. 1 compares the rule comprehensibility of three approaches.

Considering the complexity of algorithms, the decompositional approach rule extraction algorithms are tiresome as it analyses all the weights and activation values of hidden neurons for extracting rules and result in large computation cost and computation time than other two approaches. Since the eclectic approach algorithms examine the input-output relationships and also the weights and activation values of hidden neurons, this type of method may lead to obtain the high accuracy rules but their complexity is higher than the pedagogical approach. The pedagogical approach algorithms consist of portability i.e., the rules can be extracted from the network with any structure as it does not analyze the structures/weights of the network. But the network with large number of hidden nodes always has poor generalization [36]. So it is advisable to have the

network with minimum number of hidden nodes. For larger network, number of hidden nodes can be reduced by any existing pruning algorithms [37, 38].

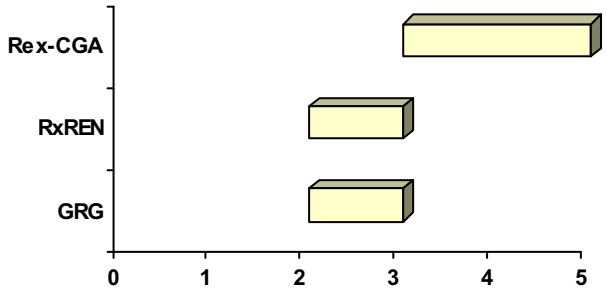

Figure 1. Rule comprehensibility comparison of three algorithms

In summary, the comparative study consistently indicate that the performance of the pedagogical approach algorithms are better than other approaches in terms of rule comprehensibility, computation cost, computation speed, portability and classification accuracy.

V. CONCLUSION

This paper discusses on various rule extraction algorithms which were developed using three different rule extraction approaches namely decompositional, pedagogical and eclectic in section II. The criterions for evaluating the extracted rules are discussed in Section III and Section IV compares the performance of the three rule extraction algorithms on three real datasets namely Wisconsin breast cancer, Pima Indian Diabetes and iris dataset.

The theoretical study in section II shows that the rules extracted using algorithms in decompositional and eclectic approaches can be more accurate than pedagogical as they analyze the internal structure of the network to interpret the network's behavior. But this internal structure analysis leads an increase in computational cost and computational time. Since the pedagogical approach doesn't analyze the internal architecture, it can be faster than both decompositional and eclectic. Also the pedagogical approach rule extraction algorithms are platform independent i.e., algorithms can be applied on any type of neural networks as it extracts rules by mapping the input output relationships. The performance analysis of three rule extraction algorithms on three experimental datasets shows that the pedagogical approach algorithm achieves better classification accuracy with higher rule comprehensibility than other algorithms.

In a nutshell, the pedagogical approach rule extraction algorithms are very effective and easy to use algorithms for extracting rules from neural networks which classifies large datasets.

REFERENCES

[1] Mantas C.J., Puche J.M., Mantas J.M., Extraction of similarity based fuzzy rules from artificial neural networks, International Journal of Approximate Reasoning, vol. 43, 2006, pp. 202-221.

[2] Brain J.Taylor, Marjorie A. Darrah, Rule extraction as a formal method for the verification and validation of neural networks, IEEE International Joint Conference on Neural Networks, vol. 5, 2005, pp. 2915 - 2920.

[3] S. Cohen, L. Rokach, O. Maimon, Decision-tree instance-space decomposition with grouped gain-ratio, Information Sciences, vol. 177, No. 17, 2007, pp. 3592-3612.

[4] K.Kaikhah, S.Doddmeti, Discovering trends in large datasets using neural network, Applied Intelligence vol. 29, 2006, pp. 51-60.

[5] H.Dam, H.A.Abbass, C.Lokan, Xin yao, Neural based learning classifier systems, IEEE Transactions on Knowledge and Data Engineering, vol. 20 , No. 1, 2008.

[6] Setiono R., Liu H., Symbolic representation of neural networks, IEEE Computer, vol. 29, No. 3, 1996, pp. 71-77.

[7] Setiono R., Extracting rules from neural networks by pruning and hidden-unit splitting, Neural Computation, vol. 9, No. 1, 1997, pp. 205-225.

[8] Amit Gupta, Sang Park, and Shwa M. Lam,Generalized Analytic Rule Extraction for Feedforward Neural Networks, IEEE Transactions on Knowledge and Data Engineering, vol. 11, No. 6, 1999.

[9] Terence A. Etchells and Paulo J. G. Lisboa, Orthogonal Search-Based Rule Extraction (OSRE) for Trained Neural Networks: A Practical and Efficient Approach, IEEE Transactions on Knowledge and Data Engineering, vol. 17, No. 2, 2006, pp. 374-384.

[10] Setiono, Baesens and Mues, Recursive neural network rule extraction for data with mixed attributes,IEEE Transactions on Knowledge and Data Engineering, vol. 19, No. 2, 2008, pp. 299-307.

[11] Emad.W.Saad, Donald C.Wunsch II, Neural network explanation using inversion, Neural networks, vol. 20, 2007, pp. 78-93.

[12] Andrews R., Diederich J., and A.B. Tickle, Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks, Knowledge based systems, vol. 8, No. 6, 1996, pp. 373-389.

[13] Setiono R. and Liu H., Understanding Neural Networks via Rule Extraction, in: Proc. of 14th International Joint Conference on Artificial Intelligence , 1995, pp. 480-485.

[14] Setiono R. and Liu H., Neurolinear: From neural networks to oblique decision rules, Neural Computing, vol. 17, No. 1, 1997, pp. 1-24 .

[15] Kim D., and Lee J., Handling continuous-valued attributes in decision tree with neural network modeling, In Proceedings of the 11th european conference on machine learning, Lecture notes in computer science, 1810, 2000, pp. 211-219.

[16] Krishnan R., Sivakumar G., Bhattacharya P., A search technique for rule extraction from trained artificial neural networks, Pattern recognition letters, vol. 20, 1999, pp. 273-280.

[17] Setiono R., and Leow W.K., FERNN: An algorithm for fast extraction of rules from neural networks. Applied Intelligence, vo. 12 No. 1-2, 2000, pp. 15-25.

[18] Tsukimoto H., Extracting rules from trained neural networks, IEEE Transactions on Neural Networks vol. 11, No. 2, 2000, pp. 377-389.

[19] Koichi Odajimaa, Yoichi Hayashi , Gong Tianxia, Rudy Setiono, Greedy rule generation from discrete data and its use in neural network rule extraction, Neural Networks vol. 21, 2008, pp. 1020-1028.

[20] Setiono R., Baesens B., Mues C., A note on knowledge discovery using neural networks and its application to credit card screening, European Journal of Operational Research vol. 192, No. 1, 2008, pp. 326-332.

[21] Lale Ozbakir, Adil Baykasoglu, Sinem Kulluk, A soft computing-based approach for integrated training and rule extraction from artificial neural networks: DIFACONNminer, Applied Soft Computing, vol. 10, No. 1, 2010, pp. 304-317.

[22] Saito K. and Nakano R. Medical diagnostic expert system based on pdp model. In Proceedings of IEEE International Conference on Neural Networks, vol. 1, 1988, pp. 255-262.

[23] Sestito S. and Dillon T.S., Automated knowledge acquitition of rules with continuously valued attributes, In proc. 12th Int. Conf. on expert systems and their applications, 1992, pp. 645-656.

[24] Thrun S.B., Extracting provably correct rules from artificial neural networks, Technical Report IAI-TR-93-5, Institut fr informatik III, University of Bonn, 1993.

[25] Pop E., Hayward R., and Diederich J., RULENEG: Extracting rules from a trained ANN by stepwise negation, Tech. Rep., QUT NRC, 1994.

[26] Craven M.W., Shavlik J.W., Using sampling and queries to extract rules from trained neural networks, in: Proceedings of the 11th International Conference on Machine Learning, San Francisco CA, 1994.

[27] Filer R., Sethi I., Austin J., A comparison between two rule extraction methods for continuous input data, In proc. of Neural information processing systems, rule extraction from trained artificial neural networks workshop, 1997, pp. 38-45.

[28] Emad.W.Saad, Donald C.Wunsch II, Neural network explanation using inversion, Neural networks vol. 20, 2007, pp. 78-93.

[29] M.Gethsiyal Augasta and T.Kathirvalavakumar, Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems, Neural Processing Letters, 2011 (online), DOI 10.1007/s11063-011-9207-8.

[30] Craven M.W., Shavlik J.W., Learning symbolic rules using artificial neural networks, in: Proceedings of the 10th International Conference on Machine Learning, San Mateo, CA, 1993, pp. 73-80.

[31] Tickle A.B., Orlowski M. and J. Diederich J., DEDEC:Decision Detection by rule extraction from neural networks, QUTNRC, 1994.

[32] Keedwell E., Narayanan A., and Savic D., Creating rules from Trained Artificial Neural Networks Using Genetic Algorithms, International Journal of Computers, Systeming Signals, vol. 1, 2000, pp. 30-42.

[33] Kaikhah K., Doddmeti S., Discovering trends in large datasets using neural network, Applied Intelligence, vol. 29, 2006, pp. 51-60.

[34] Hruschka E.R., Ebecken N.F.F., Extracting rules from multilayer perceptrons in classification problems: a clustering-based approach, Neurocomputing, vol. 70, 2006, pp. 384-397.

[35] Kahramanl H., Allahverdi N., Rule extraction from trained adaptive neural networks using artificial immune systems, Expert Systems with Applications, vol. 36, 2009, pp. 1513- 1522.

[36] Chauvin Y., Generalization performance of overtrained backpropagation networks, in: Proc. of Neural networks Euroship workshop L.B.Hlomeida and C.J.Wellekens Eds., 1990, pp. 46-55.

[37] Andries P.Engelbrecht, A new pruning heuristic based on variance analysis of sensitivity information, IEEE Transactions on Neural Networks, vol. 12, No. 6, 2001, pp. 1386-1399.

[38] Hong-Jie Xing and Bao-Gang Hu, Two phase construction of multilayer perceptrons using Information Theory, IEEE Transactions on Neural Networks vol. 20, No. 4, 2009, pp. 715- 721.