Mar. 2004, Vol.19, No.2, pp.249–253 *J. Comput. Sci. & Technol.*

Rule Extraction: Using Neural Networks or for Neural Networks?

Zhi-Hua Zhou

National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, P.R. China

E-mail: zhouzh@nju.edu.cn

Received July 21, 2003; revised October 8, 2003.

Abstract In the research of rule extraction from neural networks, *fidelity* describes how well the rules mimic the behavior of a neural network while *accuracy* describes how well the rules can be generalized. This paper identifies the *fidelity-accuracy* dilemma. It argues to distinguish *rule extraction using neural networks* and *rule extraction for neural networks* according to their different goals, where fidelity and accuracy should be excluded from the rule quality evaluation framework, respectively.

Keywords rule extraction, neural network, fidelity, accuracy, machine learning

1 Introduction

An inherent defect of neural networks is that the learned knowledge is concealed in a large amount of connections, which leads to the poor comprehensibility, i.e., poor transparency of knowledge and poor explanation ability. In order to offset this defect, developing algorithms to extract symbolic rules from trained neural networks has been a hot topic in recent years[1,2].

This paper identifies the *fidelity-accuracy dilemma* in the research of rule extraction from neural networks. It illustrates that although both fidelity and accuracy are key elements of the prevailing rule quality evaluation framework, in many cases obtaining high fidelity and high accuracy simultaneously is impossible. Moreover, this paper reveals that current research unfortunately confuses two goals, namely trying to obtain accurate and comprehensible learning systems, and trying to understand the working mechanism of neural networks. Therefore it argues to distinguish *rule extraction using neural networks* and *rule extraction for neural networks*. The former is for the first goal while the latter for the other one. Fidelity and accuracy should be excluded from the rule quality evaluation frameworks, respectively.

Although this paper does not present any new rule extraction algorithm, it is indeed about algorithms because it clarifies the goals for rule extraction algorithms and uncovers a trouble hidden in the rule quality evaluation framework, which may affect the design, evaluation, and comparison of rule extraction algorithms. The rest of this paper is organized as follows. Section 2 briefly introduces rule extraction from neural networks. Section 3 discloses the fidelity-accuracy dilemma. Section 4 explores the causation of the dilemma. Section 5 concludes the paper.

2 Rule Extraction

At the end of 1980s, Gallant[3] devised connectionist expert systems that regarded neural network as the knowledge base. In order to explain the inference process in the system, he presented a routine for extracting propositional rules from a simple network[3]. This could be viewed as the origin of the investigation on rule extraction from neural networks.

To date, one line in the development of rule extraction algorithms has been directed towards presenting the output as a set of rules using propositional logic[4−9]. A substantial parallel effort has been directed towards expressing the knowledge embodied in the neural networks using concepts drawn from fuzzy logic[10−12]. Also some algorithms have been developed to extract deterministic finite-state automata (DFA) from recurrent neural networks[13−15], and recently some researchers even have attempted to generate regression rules from neural regressors[16,17]. Although this paper mainly focuses on algorithms for extracting propositional rules, it is worth noting that most discussions may also apply to the extraction of other kinds of rules.

*Note

The work was supported by the National Outstanding Youth Foundation of China under Grant No.60325237 and the National Natural Science Foundation of China under Grant No.60273033.

According to the taxonomy presented by Andrews *et al.*[1], rule extraction algorithms can be roughly classified into three categories, namely the *decompositional*, *pedagogical*, and *eclectic* algorithms. The decompositional algorithms extract rules from each unit in a neural network and then aggregate them. Representatives of this category include Subset[4], COMBO[7], RX[8], etc. The pedagogical algorithms regard the trained neural network as an *opaque* and aim to extract rules that map inputs directly into outputs. Representatives of this category include VIA[5], TREPAN[6], STARE[9], *etc.* The *eclectic* algorithms incorporate elements of both decompositional and pedagogical ones. Representatives of this category include DEDEC[18] and the algorithm proposed by [19]. It is worth mentioning that Tickle *et al.*[2] extended the taxonomy by appending the fourth category, namely *compositional* algorithms. These algorithms are not strictly decompositional because they do not extract rules from individual units with subsequent aggregation to form a global relationship, nor do them fit into the eclectic category because there is no aspect that fits the pedagogical profile. Representatives of this category mainly include algorithms for extracting DFA from recurrent neural networks[13−15].

A salient theoretical discovery in this area is that, in many cases, the computational complexity of extracting rules from trained neural networks and the complexity of extracting rules directly from the data are both NP-hard[20]. It is also worth mentioning that Roy[21] wisely disclosed the conflict between the idea of rule extraction and traditional connectionism. In detail, the idea of rule extraction from a neural network involves certain procedures, specifically the reading of parameters from a network, which is not allowed by traditional connectionist framework that these neural networks are based on. Fortunately, Roy[21] indicated that such a conflict could be resolved by introducing some control theoretic paradigm that has been supported by new evidence from neuroscience about the role of neuromodulators and neurotransmitters in the brain.

Note that this section just briefly introduces the area of rule extraction from neural networks, which does not want to provide a thorough review. More references can be found in some good reviews[1,2,22].

3 The Fidelity-Accuracy Dilemma

Andrews *et al.*[1] presented a framework, namely FACC, for evaluating the quality of the rules extracted from neural networks, which is the prevailing rule quality evaluation framework in this area until now. The FACC framework comprises four criteria, namely *fidelity*, *accuracy*, *consistency*, and *comprehensibility*. Fidelity describes how well the rules mimic the behavior of a neural network, which is usually defined as the percentage of test examples for which the classification made by the rules agrees with the neural network counterpart. Accuracy describes how well the rules can be generalized, which is usually defined as the percentage of test examples that are correctly classified by the rules. Consistency is given if the rules extracted under different training sessions produce the same classifications of test examples. Comprehensibility is determined by measuring the number of rules and the number of antecedents per rule. In principle, all these quality measures should be pursued during the rule extraction process, but for simplifying the discussion, in this paper we mainly focus on fidelity and accuracy.

In a recent paper, Zhou *et al.*[23] presented a pedagogical algorithm for extracting prepositional rules from a complicated neural network system. With different configurations the presented algorithm can extract rules with high fidelity but moderate accuracy or high accuracy but moderate fidelity, as summarized in Table 1. A particular interesting issue that has not been addressed[23] is: which configuration should we prefer?

Table 1. Summary of Tables 6–7 of [23]

	fidelity	*accuracy*
config-1	high	moderate
config-2	moderate	high

This question places us to an embarrassing situation: to sacrifice the fidelity, or to sacrifice the accuracy. In fact, pursuing high fidelity and high accuracy may not be possible in certain situations, although this has not been recognized before. Here we call the situation as the fidelity-accuracy dilemma.

More formally, let X denote the test set, $h(x)$ denote the desired function, $f_N(x)$ denote the function implemented by the trained neural network, and $f_R(x)$ denote the function implemented by the rules extracted from the trained network. The fidelity and the accuracy of the rules can be defined as (1) and (2), respectively.

$$fidelity_R = 1 - \text{Prob}\{x \in X | f_R(x) \neq f_N(x)\} \quad (1)$$

$$accuracy_R = 1 - \text{Prob}\{x \in X | f_R(x) \neq h(x)\} \quad (2)$$

Suppose a new test example, i.e., t is appended to X, given $f_N(t) \neq h(t)$, and the rules can be modified for t. Now, if $f_R(t)$ equals $f_N(t)$, then the accuracy is deteriorated; but if $f_R(t)$ equals $h(t)$, then the fidelity is deteriorated. This illustrates that improving the accuracy does not necessarily lead to the improvement of the fidelity, and vice versa. In fact, a rule extraction algorithm may implicitly detect an error made by the trained neural network during the rule extraction process. If it tries to mimic the error then it loses the accuracy, otherwise it loses the fidelity.

It is interesting to note that many researchers observed that the extracted rules may be generalized better than the trained neural network from which the rules were extracted[23−28]. If they attempt to improve the fidelity of the rules further, they might have great chances to find that the accuracy of the rules is deteriorated unfortunately. This is because in order to pursue 100% fidelity, the accuracy of the rules must be decreased to the level of the accuracy of the trained neural network.

4 The Role of Fidelity

In order to explore the causation of the fidelity-accuracy dilemma, we should re-examine the goal of rule extraction from neural networks, and the role that fidelity plays in the FACC framework ①.

It is well known that in many cases neural networks and other learning techniques such as decision trees can achieve similar performance, and in general there is no technique which is consistently superior to another according to the No Free Lunch theorem[29]. However, it is also well known that some specific problems may be more suitable to neural networks, while others may be more suitable to other learning techniques[30]. This is the reason why neural networks have been widely investigated and applied.

As Andrews *et al.*[1] indicated, the motivation of rule extraction from neural networks is to facilitate neural networks with some form of explanation capability. In other words, since neural networks may be more accurate ② than other learning techniques for some specific problems, its potential can be better realized if its comprehensibility is enhanced. But if neural network is not so accurate as some other learning techniques for a specific problem, it is the alternative learning techniques instead of neural network that should be used. Therefore it is evident that the principal goal of rule extraction from neural networks is to generate accurate and comprehensible learning systems, where neural network is only one tool for achieving this goal. We call this scenario as rule extraction using neural networks.

Now if we examine the FACC framework, we may be astonished to find that the role of fidelity is to require the extracted rules to faithfully exhibit the behavior of the trained neural network, which has nothing to do with the goal of rule extraction using neural networks. However, since both fidelity and accuracy have been emphasized in the FACC framework, most studies in this area have attempted to extract rules with both high fidelity and high accuracy, although as indicated in Section 3, this is impossible in certain situations. Even more, some researches focused on only fidelity, or claimed that 100% or arbitrarily high fidelity had been achieved[19,31−33]. Such a stress on fidelity has also influenced some studies on extracting DFA from recurrent neural networks or extracting regression rules[17,34].

Based on above analysis, it is evident that fidelity should be excluded from the rule quality evaluation framework for rule extraction using neural networks. Moreover, the question raised in Section 3 could be easily answered: the configuration with high accuracy is preferable because here the goal is to obtain accurate and comprehensive learning systems.

Then another question arises: is fidelity useless at all?

In answering this question, we must be very cautious. Although fidelity seems useless in achieving the goal of developing accurate and comprehensible learning systems, it might be useful for other purposes. In fact, rule extraction from neural networks may have a secondary goal, that is, to understand the working mechanism of trained neural networks. In order to achieve this goal, the rules should replicate the behavior of the trained neural network from which they were extracted as

①Heuristically, we believe that the fidelity-accuracy dilemma is related to the amount of available training examples, the complexity of the learning task, and the complexity of the trained neural networks. However, in this paper we only focus on the relationship between the dilemma and the FACC framework, leaving the exploration of the nature of the dilemma as future work.

②In fact, there are several other factors, such as the ease of knowledge representation, which may affect our decision of whether to use neural network or not. But for simplifying the discussion, here we mainly focus on accuracy.

faithfully as possible. We call this scenario as rule extraction for neural networks.

It is obvious that in this scenario fidelity is a key criterion in evaluating the rule quality. Moreover, accuracy is useless now because as far as the extracted rules faithfully reproduce the trained neural network, we do not care whether they generalize well or not. It is even more interesting that having a higher extracted rule accuracy over the original network is not desirable now, because these rules cannot tell us how the trained network actually works.

5 Concluding Remarks

This paper presents a critique to FACC, which is the prevailing rule quality evaluation framework in the area of rule extraction from neural networks. This paper shows that two different goals of rule extraction have been confused, which causes the fidelity-accuracy dilemma. It argues to distinguish rule extraction using neural networks and rule extraction for neural networks according to their differing goals. Furthermore, it argues that the ACC (accuracy, consistency, and comprehensibility) instead of the FACC framework should be used for rule extraction using neural networks, while the FCC (fidelity, consistency, and comprehensibility) instead of the FACC framework should be used for rule extraction for neural networks.

Acknowledgements This paper was written when the author was visiting the Academy of Mathematics and Systems Sciences, CAS, China.

References

[1] Andrews R, Diederich J, Tickle A B. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 1995, 8(6): 373–389.

[2] Tickle A B, Andrews R, Golea M *et al.* The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Trans. Neural Networks*, 1998, 9(6): 1057–1067.

[3] Gallant S I. Connectionist expert systems. *Communications of the ACM*, 1988, 31(2): 152–169.

[4] Fu L. Rule learning by searching on adapted nets. In *Proc. the 9th National Conference on Artificial Intelligence*, Anaheim, CA, 1991, pp.590–595.

[5] Thrun S. Extracting rules from artificial neural networks with distributed representations. In *Advances in Neural Information Processing Systems 7*, Tesauro G, Touretzky D, Leen T (Eds.), Cambridge, MA, MIT Press, 1995, pp.505–512.

[6] Craven M W, Shavlik J W. Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems 8*, Touretzky D, Mozer M, Hasselmo M (Eds.), Cambridge, MA, MIT Press, 1996, pp.24–30.

[7] Krishnan R. A systematic method for decompositional rule extraction from neural networks. In *Proc. the NIPS'96 Workshop on Rule Extraction from Trained Artificial Neural Networks*, Queensland, Australia, 1997, pp.38–45.

[8] Setiono R. Extracting rules from neural networks by pruning and hidden-unit splitting. *Neural Computation*, 1997, 9(1): 205–225.

[9] Zhou Z H, Chen S F, Chen Z Q. A statistics based approach for extracting priority rules from trained neural networks. In *Proc. the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Como, Italy, 2000, 3: 401–406.

[10] Masuoka R, Watanabe N, Kawamura A *et al.* Neurofuzzy systems — Fuzzy inference using a structured neural network. In *Proc. the International Conference on Fuzzy Logic and Neural Networks*, Iizuka, Japan, 1990, pp.173–177.

[11] Mitra S. Fuzzy MLP based expert system for medical diagnosis. *Fuzzy Sets and Systems*, 1994, 65(2-3): 285–296.

[12] Castro J L, Mantas C J, Benitez J M. Interpretation of artificial neural networks by means of fuzzy rules. *IEEE Trans. Neural Networks*, 2002, 13(1): 101–116.

[13] Giles C L, Miller C B, Chen D *et al.* Learning and extracting finite state automata with second-order recurrent neural networks. *Neural Computation*, 1992, 4(3): 393–405.

[14] Omlin C W, Giles C L, Miller C B. Heuristics for the extraction of rules from discrete time recurrent neural networks. In *Proc. the International Joint Conference on Neural Networks*, Baltimore, MD, 1992, Vol.1, pp.33–38.

[15] Giles C L, Omlin C W. Extraction, insertion, and refinement of symbolic rules in dynamically driven recurrent networks. *Connection Science*, 1993, 5(3-4): 307–328.

[16] Saito K, Nakano R. Extracting regression rules from neural networks. *Neural Networks*, 2002, 15(10): 1279–1288.

[17] Setiono R, Leow W K, Zurada J M. Extraction of rules from artificial neural networks for nonlinear regression. *IEEE Trans. Neural Networks*, 2002, 13(3): 564–577.

[18] Tickle A B, Orlowski M, Diederich J. DEDEC: A methodology for extracting rule from trained artificial neural networks. In *Proc. the AISB'96 Workshop on Rule Extraction from Trained Neural Networks*, Brighton, UK, 1996, pp.90–102.

[19] Craven M W, Shavlik J W. Using sampling and queries to extract rules from trained neural networks. In *Proc. the 11th Int. Conf. Machine Learning*, New Brunswick, NJ, 1994, pp.37–45.

[20] Golea M. On the complexity of rule extraction from neural networks and network querying. In *Proc. the AISB'96 Workshop on Rule Extraction from Trained Neural Networks*, Brighton, UK, 1996, pp.51–59.

[21] Roy A. On connectionism, rule extraction, and brain-like learning. *IEEE Trans. Fuzzy Systems*, 2000, 8(2): 222–227.

[22] Duch W, Adamczak R, Grabczewski K. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Trans. Neural Networks*, 2001, 12(2): 277–306.

[23] Zhou Z H, Jiang Y, Chen S F. Extracting symbolic rules from trained neural network ensembles. *AI Communications*, 2003, 16(1): 3–15.

[24] Towell G, Shavlik J. The extraction of refined rules from knowledge based neural networks. *Machine Learning*, 1993, 13(1): 71–101.

[25] Fu L. Rule generation from neural networks. *IEEE Trans. Systems, Man and Cybernetics*, 1994, 24(8): 1114–1124.

[26] Craven M W, Shavlik J W. Extracting comprehensible concept representations from trained neural networks. In *Working Notes on the IJCAI'95 Workshop on Comprehensibility in Machine Learning*, Montreal, Canada, 1995, pp.61–75.

[27] Taha I A, Ghosh J. Symbolic interpretation of artificial neural networks. *IEEE Trans. Knowledge and Data Engineering*, 1999, 11(3): 448–463.

[28] Setiono R. Extracting M-of-N rules from trained neural networks. *IEEE Trans. Neural Networks*, 2000, 11(2): 512–519.

[29] Wolpert D H, Macready W G. No free lunch theorems for optimization. *IEEE Trans. Evolutionary Computation*, 1997, 1(1): 67–82.

[30] Quinlan J R. Comparing connectionist and symbolic learning methods. In *Computational Learning Theory and Natural Learning Systems*, Rivest R L (Ed.), Vol.1, Cambridge, MA, MIT Press, 1994, pp.445–456.

[31] Chalup S, Hayward R, Diederich J. Rule extraction from artificial neural networks trained on elementary number classification tasks. In *Proc. the 9th Australian Conference on Neural Networks*, Brisbane, Australia, 1998, pp.265–270.

[32] Maire F. Rule-extraction by backpropagation of polyhedra. *Neural Networks*, 1999, 12(4-5): 717–725.

[33] Bologna G. Rule extraction from a multi layer perceptron with staircase activation functions. In *Proc. the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Como, Italy, 2000, 3: 419–424.

[34] Vahed A, Omlin C W. Rule extraction from recurrent neural networks using a symbolic machine learning algorithm. In *Proc. the 6th International Conference on Neural Information Processing*, Dunedin, New Zealand, 1999, pp.712–717.