Knowledge Extracted from Trained Neural Networks – What's Next?

J.T. Yao

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
E-mail: jtyao@cs.uregina.ca

ABSTRACT

One of the major drawbacks or challenges of neural network models is that these models can not explain what they have done. Extracting rules from trained neural networks is one of the solutions for understanding the networks. However, what we should do with these extracted rules remains a research question. This paper tries to address issues on effectively and efficiently utilizing extracted rules or knowledge.

Keywords: Neural networks, rule extraction, descriptive neural networks.

1. INTRODUCTION

Forecasting future events are very important to our daily life. Much effort has been made to research on prediction and classification. Statistics and neural networks offer the best fit to the data and are a satisfactory solution for such problems. Formulation of understandable rules derived from data analysis is more important than the forecasting models themselves.[4]

Neural networks are computer software that tries to emulate biological neural networks. A neural network may act as a learning system made up of simple units configured in a highly interconnected network. It is aimed to solve a problem that is hard to find a well formulated algorithmic solution. For instance, we want to find the underlying rules or structures from existing large amount of data or examples. The major difference of neural networks with traditional computing is that neural networks are based on the parallel architecture of animal brains. It has been reported in literature that neural networks are good candidates for solving classification and forecasting problems.[5, 9, 18, 27] Despite the success of neural networks in different applications, one of the major concerns preventing further development of neural networks is that there is no explanation of the mechanism inside the models.

Extracting rules from trained neural networks is one of the solutions for understanding the networks. Many researchers have shown that useful rules could be obtained from trained neural networks.[7, 16, 19] However, what should we do with these extracted rules remains a research question. It does not take full advantage of the neural network technique by just using extracted rules. The majority of the rules extracted do not have nonlinearity as in neural networks. In addition, there are other better rule extraction approaches other than neural networks, e.g. data mining in general. It was argued that by incorporating extracted rules with a neural network to reconstruct a descriptive neural network may help us to gain more understanding of neural network mechanisms and make more accurate forecastings.[25] It was reported that many researchers confuse two different goals of studies of rule extraction from neural networks.[30] The first goal is to obtain accurate and comprehensible learning systems, and the second goal is understanding the working mechanism of neural networks. To distinguish rule extraction by using neural networks and rule extraction for neural networks understanding may be of important. Knowledge based descriptive neural networks may fulfil these two goals under one umbrella.[25]

This paper tries to address issues on effectively and efficiently utilizing extracted rules or knowledge from neural networks. It is organized as follows. The next section will summarize neural network and data mining techniques. A section that introduces neural network rule extraction follows. Some possible mechanisms that utilize extracted rules are discussed in Section 4. Finally, we conclude this paper.

2. NEURAL NETWORKS

Artificial neural networks have the ability to learn from examples and exhibit some capability for generalization beyond the training data. They are very useful when patterns or trends that are too complex to be noticed by either humans or other computer software. Indeed, they have been used in such diverse applications as pattern recognition,[15] medical diagnosis,[14] foreign exchange prediction,[27] stock market assessment and prediction,[28] and many more. Many researchers have suggested that neural networks can serve as alternative and novel tools in business, e.g., in forecasting financial time series.[24] In addition, neural networks have been mathematically proved to be universal approximators of functions and their derivatives.[22] Hence the potential benefits to various applications may be unlimited.

Neural networks can be classified as one of soft computing techniques. Soft computing is a collection of techniques spanning many fields that fall under various categories in computational intelligence.[29] Soft computing methodologies including neural networks, genetic algorithms, fuzzy sets, rough sets and wavelet are widely applied in data mining and the knowledge discovery process.[10]

Data mining or knowledge discovery in databases is defined as the the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[6] Generally speaking, data mining approaches can be categorized into two classes: descriptive data mining and predictive data mining. Descriptive data mining explores interesting and unknown patterns that describe the data. Predictive data mining forecasts the behavior of events based on the model obtained on available data. In other words, the former is to discover knowledge, the later uses the knowledge discovered. Neural networks are good predictive models. However, due to the black-box nature of neural networks, they are sometimes not classified as a data mining tool for discovering interesting and understandable patterns according to data mining definition.[6] There is no information or knowledge embedded in trained neural networks that can be verified or interpreted by human beings. If one cannot reason the results obtained from a neural network on a logical level, it hard for a user to use this model in critical application areas. Extracting rules from trained neural networks in the aim of explaining mechanism of neural networks is one of the solutions to resolve the dilemma.

3. EXTRACT RULES FROM NEURAL NETWORKS

As mentioned above, being unable to explain the knowledge embedded in trained neural networks is one of the major drawbacks of this technology. Much attention has been paid to solve this problem by extracting rules from trained neural networks. Many researchers have been working on extracting rules from trained networks, as reasoning with logical rules is more acceptable to users than black box systems.[1] Setiono's work[16] is one of the examples to address this problem and open the black box of neural networks.

The reasons for extracting rules for a neural network may vary according to Wall *et. al.* by summarizing others work.[21]

- Extracting rules is trying to find user explanations of the system. This will help a novice user to gain insight into a problem by understanding the internal logic of the system.
- A safety critical system, such in health care domains or diagnostic control of dangerous systems, must only produce results that are provably correct. The system has to disregard some results that are wrong or potentially wrong under certain circumstances, rule extraction can help to identify these cases.
- Formal methods such as Vienna Development Method (VDM)[23] aim to develop software that is verified of their operation before deployment. That is, the only reliable system is the one that has been proved correct mathematically. The current rule extraction techniques do not have the ability to prove that a neural network will act according to a predetermined specification. However, partial, even whole, understanding of the operation of the rules is a good start to achieve this goal.
- When the data is insufficient, neural network generalization ability may get weak. A rule system with good interpretability improves neural network generalization ability.
- Nonlinearity is the key advantage of neural networks. From a data mining point of view, previous unknown complex nonlinear relationships from the given data set may be exposed to the user by using rule extraction techniques.

According to the taxonomy of Tickle *et al.*,[19] neural network rule extraction techniques may be classified into five dimensions.

- The expressive power of the extracted rules. It is mainly for the format or type of rules, e.g., symbolic, propositional, boolean, probabilistic, fuzzy, or first-order-logic rules.
- The quality of the extracted rules. The quality can be measured by accuracy, fidelity, consistency and comprehensibility. The accuracy reflects the forecastibity of the rules. The fidelity represents the ability to mimic the behavior of the neural networks. It was argued that it may be hard to achieve high level of accuracy and fidelity at the same time for certain applications. Thus only one of them might be chosen for different goals.[30] The consistency is about the robustness of the rules across different unforseen data sets. The comprehensibility is the measure of the size of the rule set and the rules themselves.[1]
- The translucency of the view taken within the rule extraction technique of the underlying network units. Pedagogical techniques do not analyze detailed characteristics of the neural networks. They only extract global relationship between inputs and outputs without. Decompositional group of techniques examine individual (level of) neurons in order to aggregate to form a global relationship. Eclectic techniques is in between pedagogical and decompositional.
- The complexity of the algorithms. The portability to network architectures and training regimes. Some of the techniques are only applied to certain network architecture or certain applications.
- A sixth dimension, the treatment of linguistic variables has been considered by Duch *et al.*[3] Some methods work only with binary variables, while the others may work with discrete or continuous variables.

The fifth criterion of taxonomy of Tickle *et al.*[19] is to measure the generalization ability of the technique. This dimension can be extended to applications as well. For instance, in the area of financial forecasting with neural networks the rules should be as simple as possible from a practitioner's point of view.[25] In a more critical application area, accuracy may be set as the first priority. In any case, the rules extracted from neural networks should be easily interpreted and transferred into actions by users. To adopt or propose a suitable tool is not an easy job for most applications. Almost all factors should be considered. As posited by Duch,[3] most research papers on the rule extraction are usually limited to the description of new algorithms. Only a partial solution to the problem of knowledge extraction from data are presented. However, control of the tradeoff between comprehensibility and accuracy, optimization of the linguistic variables and final rules, and estimation of the reliability of rules are almost never discussed. To resolve this dilemma, one may start from a rough, low accuracy, simple description of the data. More accurate, complex description rules may be introduced later. Neural network rule extraction methods may serve, at least as initial rules, but that should not be the end of the story.[3]

There are many types of rule formats, prepositional IF ... THEN rules may be of the simplest type. It can be described as

$$\text{IF } \; x_1 \in X_1 \wedge x_2 \in X_2 ... x_n \in X_n \quad \text{THEN } \; Class = C_k. \tag{1}$$

X_i can be a set of symbolic values, discrete numerical values or intervals for continuous features. Equation 1 forms a crisp logic rule. It is good if all conditions meet. However, it is unwontedly untrue in majority of time. M-of-N rules that seek for a partial solution are also popular rule format.[16] The general idea is that it is true if at least M out of a set of N features are presented.

4. EFFICIENT AND EFFECTIVE USAGES OF RULES EXTRACTED FROM TRAINED NEURAL NETWORKS

The predictive power of neural networks has been shown in literature. Although prediction of stock changes is important in financial markets, understanding the factors that result in the changes is more important. The rules like "When the 10-day moving average crosses above the 30-day moving average and both moving averages are in an upward direction it is the time to buy" and When the 10-day moving average crosses below the 30-day moving average and both moving averages are directed downward it is time to sell" are more understandable than a black box neural network model. Traders would like to use rules rather than the neural network model.

Combining predictive models with *a priori* knowledge about the problem is usually difficult.[4] In most of cases, the big challenge is that we have no *a priori* knowledge. The data mining tasks are somehow aimless. However, "Let the data speak for themselves" is ill-suited for forecasting according to Armstrong forecasting principles.[2] There is no way to test or control the models in the ears of the future space that are far from the training data. It may not be acceptable by practitioners if only statistic regression or neural network models are used as forecasting tools, especially in safety critical domains like medical and aviation industrial. The aim of this study is to identify possible ways to combine extracted rules with neural networks. In this way, we can still keep the advantages of neural networks. In addition understandable and descriptive human knowledge is embedded into the predictive model. The models can then be considered as either predictive and descriptive models. They are more acceptable to practitioners. The descriptive neural network models,[25] though mainly for financial prediction, can be viewed as one step towards our goal.

4.1. Descriptive Neural Networks

A descriptive neural network (DNN) is a neural network embedded with business rules that have been discovered from previously trained networks.[25] The architecture of DNN is not only decided by training examples but also by hidden rules extracted from trained networks. The DNN system to traditional neural network is similar to econometrics to regression analysis. It may not be acceptable by practitioners if only regression or neural network models are used as forecasting tools. Econometrics is the technique using statistical analysis combined with economic theory to analyze economic data. It presents more economic knowledge than a single mathematical formula and thus more popular and acceptable to practitioners. In fact, it is more accurate than regression models.

There are some ways to utilize rules obtained from trained neural networks. First, the extracted rules can be treated as an add-on to neural network models. The main aim is not to actually utilize these rules. They are only served for demonstrative or explanatory purpose. This is especially true for rules obtained by global rule extraction methods. The second approach is to adopt the rules extracted as *a priori* knowledge. To incorporate these rule into neural networks or utilize them for construction of neural networks. If the knowledge of an input factor has great influence to the output is known. Instead of treating this input equal with other inputs, we may connect it directly to the model's output or at least with a shot path. When training a neural network, weights are generally initialized randomly according to predetermined heuristics. In this case, we can set weights according the known factors. The learning heuristic may also be adopted from the knowledge obtained.

There are three steps involved in the construction of the proposed DNN networks. The first step is to build a neural network model. A neural network construction system as presented in [25] could be used. The second step is to extract rules from trained neural networks. The techniques discussed in previous sections can be used.

In the third step, rules extracted in the previous step are incorporated into the network generated by the neural network construction system to form a descriptive neural network. Most researchers extract if-then type association rules, as they are more understandable for humans than other representations. The rules created from neural networks can then be converted to decision trees or used in other expert systems. The uses of rules here have limited neural network's ability to model nonlinear data. DNN is an artificial neural network with descriptions of the domain knowledge of applied area, so that not only predictions can be made but also the reasons for the predictions can be explained. The rules are used in construction of neural networks in terms of the architecture and weights. The weights to or from the most influential factors are set the highest in the retraining. Some unimportant factors could be passed. For instance, if we find that the daily movement has the least influence to long term forecasting, we could eliminate the node or the neural network that are served as the daily movement component.

Some issues in construction of DNN include knowledge based management, architecture enhancement, rule measurement criteria, threshold adjustment, fuzzy representation, etc.

4.2. Pruned Neural Networks

Pruning is a process to eliminate unnecessary and unimportant connections or nodes from a neural network. It is considered as one of the steps in some rule extraction techniques. For instance, the NeuroRule algorithm[17] uses symbolic representations to make each prediction explicit and understandable. It is claimed that the neural network can be interpreted by the rules which, in general, preserve network accuracy and explain the prediction process. The main idea is to compare with the weights to decide the importance of the neural network connections. Irrelevant connections are removed (pruned) if it does not decrease the predictive accuracy. Rules are then extracted from the pruned network.

Sensitivity analysis is another way that tries to simplify the complex and incomprehensible neural network.[26] Weights or inputs are analyzed in order to find importance of correlations between inputs and outputs. If the rules can be extracted from simplified neural networks by pruning or sensitivity analysis, explanation or understanding can be obtained easily. For instance, if there are only two inputs left in a network and we understand that one input has a larger connection weight to the output.

For example, the rule with the format of

$$\begin{array}{llll} \text{IF} & \alpha_2 = 2, & \text{THEN} & \text{EI}, \\ \text{ELSE} & \text{IF} \quad \alpha_1 = 2 & \text{THEN} & \text{IE}, \\ \text{ELSE} & \text{N} & & \end{array}$$

was extracted in the illustrative example in.[16] It represent a class of

$$(\alpha_1, \alpha_2) = \begin{cases} (2,1) & \text{Class IE} \\ (1,2) & \text{Class EI} \\ (2,2) & \text{Class EI} \\ (1,1) & \text{Class N} \end{cases}$$

where $\alpha_i = 1$ stands for the interval range of [-1,0.96], and $\alpha_i = 2$ [0.96,1].

4.3. Fuzzy Neural Networks

Fuzzy neural networks[8, 11] can be considered as the simplest descriptive neural networks. They provide some fuzzy explanations to the neural network mechanism. They can be viewed as a fuzzy system that uses a learning algorithm derived from neural networks to determine fuzzy rules. Linear functions such as

$$\hat{y} = f_1(x_1) + f_2(x_2) + ... + f_m(x_m) = \sum_{i=1}^{m} f_i(x_i)$$

could be used.[11] Each f_i in fact is represented by the following rules:

$$\begin{array}{lllll} R^1: & \text{IF} & x_i \text{ is } A_{i1} & \text{THEN} & y_{i1} = w_{i1} \\ R^2: & \text{IF} & x_i \text{ is } A_{i2} & \text{THEN} & y_{i2} = w_{i2} \\ & & ... & & \\ R^n: & \text{IF} & x_i \text{ is } A_{in} & \text{THEN} & y_{in} = w_{in} \end{array}$$

4.4. Ensemble Neural Networks

A neural network ensemble is a set of separately trained neural networks that are combined to form one unified prediction model.[12] Each trained neural network in the collection of ensembles can serve a different rule in modeling. Suppose we use three neural networks in financial forecasting. They are used to forecast different movements such as primary trend, secondary movements and day fluctuations as in Dow Jones theory. These networks are served as committee members in the forecasting and decision making processes. It is useful when there is disagreement among them.

There are two types of techniques, namely, Bagging and Boosting. The former creates individuals for its ensemble by training each model on a random redistribution of the training set.[13] In this case, some original examples may be repeated in the training set while others may be left out. The later chooses training examples by performance. Examples that were incorrectly predicted are chosen more often. It is hoped that better prediction can be made for those poor performed. The basic principle is that each neural network may generate different outputs. In fact, if the outputs are identical, there is no need to form an ensemble. Each individual may represent an aspect of the truth which in turn can form a understandable rule. In this way a more accurate and explicable classification or forecasting model is generated.

5. CONCLUSION

Neural networks are widely used in applications such as forecasting, pattern recognition, and classification. Although research has shown that neural networks are more effective and accurate in many areas than traditional statistical models, industrial representatives are reluctant to invest in such a technology due to the lack of explanation of its mechanism of modeling. Extracting rules from trained neural networks is one of the solutions to this problem. However, the use rules extracted from neural network does not take full advantage of what a neural network gas to offer. It is argued in this paper, there are way to utilize rules obtained from trained neural networks effectively and efficiently. Issues of incorporating discovered rules into neural networks are discussed.

REFERENCES

1. R. Andrews, J. Diederich, and A. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, *Knowledge Based Systems*, **8**(6), 373-389, 1995.
2. J. S. Armstrong, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer Academic Publishers, 2001.
3. W. Duch, R. Adamczak and K. Grabczewski, A new methodology of extraction, optimization and application of crisp and fuzzy logical rules, *IEEE Trans Neural Networks*, **12**(2), 277-306, 2001.
4. W. Duch, R. Setiono, and J.M. Źurada, Computational Intelligence Methods for Rule-based Data Understanding, *Proceedings of IEEE*, **92(**5), 771-805, 2004.
5. M. Duhoux, J. A. K. Suykens, B. De Moor, and J. Vandewalle, Improved Long-Term Temperature Prediction by Chaining of Neural Networks, *International Journal of Neural Systems*, **11**(1), 1-10, 2001.
6. U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, **17**(3), 37-54, 1996.
7. L. M. Fu, Rule Generation from Neural Networks, *IEEE transactions on systems, man and cybernetics*, **28**(8), 1114-1124, 1994.
8. N. K. Kasabov, Learning fuzzy rules and approximate reasoning in fuzzy neural networks and hybrid systems, *Fuzzy Sets and Systems*, **82**(2), 135-149, 1996.
9. S. Lawrence, I. Burns, A. Back, A. Tsoi and C. Giles, Neural Networks Classification and Prior Class Probabilities, in Tricks of the trade, Lecture Notes in Computer Science State-of-the-Art Surveys, Springer Verlag, 1998, pp299-314.
10. S. Mitra, S. K. Pal and P. Mitra, Data mining in soft computing framework: A survey, *IEEE Transactions on Neural Networks*, **13**, 3-14, 2002.
11. S.-K. Oh, W. Pedrycz, H.S. Park, Hybrid Identification in Fuzzy-Neaural Networks, *Fuzzy Sets and Systems*,**138**(2), 399-426, 2003.
12. D. Opitz and J. Shavlik, Actively searching for an effective neural-network ensemble, *Connection Science*, **8**, 337-353, 1996.
13. D. Opitz and R. Maclin, Popular Ensemble Methods: An Empirical Study, *Journal of Artificial Intelligence Research*, **11**, 169-198, 1999.
14. E.B. Reategui, J.A. Campbell and B.F. Leao, Combining a neural network with case-based reasoning in a diagnostic system *Artificial Inteligence in Medicine*, **9**(1), 5-27, 1997.
15. B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.
16. R. Setiono, Extracting M-of-N Rules From Trained Neural Networks, *IEEE Transactions on Neural Networks*, **11**(2), 512-519, 2000.
17. R. Setiono and H. Liu, Symbolic representation of neural networks, *IEEE Computers*, **29**(3), 71-77, 1996.
18. C.L. Tan, T. S. Quah, H.H. Teh, An artificial neural network that models human decision making, *Computer* **29**(3), 64-70, 1996
19. A. Tickle, R. Andrews, M. Golea and J. Diederich, The Truth Will Come To Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks, *IEEE Trans Neural Networks*, **9**(6), 1057-1068, 1998.
20. G. G. Towell, Jude W. Shavlik, Extracting Refined Rules from Knowledge-Base Neural Networks *Machine Learning*, **13**(1),71-101, 1992.

21. R. Wall and P. Cunningham, (2000) Exploring the Potential for Rule Extraction from Ensembles of Neural Networks, *11th Irish Conference on Artificial Intelligence & Cognitive Science*, 2000, pp52-68.
22. H. White, Learning in Artificial Neural Networks: A Statistical Perspective, *Neural Computation*, **1**, 425-465, 1989.
23. M. Woodman and B. Heal, *Introduction to VDM*, McGraw-Hill, 1993.
24. B. K. Wong, T. A. Bodnovich, and Y Selvi, Neural Network Applications in Business: a Review and Analysis of the Literature (1988-1995), *Decision Support Systems*, **19**(4),301-320, 1998.
25. J. T. Yao, Knowledge Based Descriptive Neural Networks, *Proceedings of the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, May 26-29, 2003, Chongqing, China, Lecture Notes in Computer Science 2639, pp430-436.
26. J. T. Yao, Sensitivity Analysis for Data Mining, *Proceedings of The 22nd International Conference of NAFIPS*, July 24-26, Chicago, USA, 2003, pp272-277.
27. J.T Yao, Y.L. Li, C. L. Tan, Forecasting the Exchange Rates of CHF vs USD Using Neural Networks, *Journal of Computational Intelligence in Finance*, **5**(2), 7-13, 1997.
28. J.T Yao, C. L. Tan and H.-L. Poh, Neural Networks for Technical Analysis: A Study on KLCI, *International Journal of Theoretical and Applied Finance*, **2**(2), 221-241, 1999
29. L. A. Zadeh, The roles of fuzzy logic and soft computing in the conception, design and deployment of intelligent systems, in H. S. Nwana and N. Azarmi (Eds), *Software Agents and Soft Computing: Towards Enhancing Machine Intelligence, Concepts and Applications*, Springer Verlag, 1997, pp183-190.
30. Z.-H. Zhou, Rule extraction: using neural networks or for neural networks? *Journal of Computer Science and Technology*, **19**(2), 249 - 253, 2004.