Covenant Journal of Informatics & Communication Technology Vol. 4 No. 2, Dec, 2016

An Open Access Journal, Available Online

Analysis of Modified Rule Extraction Algorithm and Internal Representation of Neural Network

Vinita Srivastava[1] & Chitra Dhawale[2],

[1]Department of I.I.C.C. , R.T.M. Nagpur University, Nagpur India.
[1]vinni.sara12@gmail.com
[2]Department of Computer Science, Amravati India.
[2] cadhawale@rediffmail.com

Abstract: Classification and Rule extraction is an important application of Artificial Neural Network. To extract fewer rules from multilayer feed forward neural network has been a research area. The internal representation of the network is augmented by a distance term to extract fewer rules from the feed forward neural network and experimented on five datasets. Understanding affect of different factors of the dataset and network on extraction of a number of rules from the network can reveal important pieces of information which may help researchers to enhance the rule extraction process. This work investigates the internal behavior of neural network in rule extraction process on five different dataset.

Keywords: Rule extraction, Feed Forward Neural Network, Hidden units, Activation value, Hidden neurons.

1. Introduction

Classification and pattern recognition is one of the important application of Artificail Neural Network [Kamruzzan et al. 2011][Sheng and Qi 2011]. Classification rules, extracted by efficient rule extraction algorithm helps in decision process but concepts learned by neural networks are difficult to understand because they are represented using large assemblages of real valued parameters [Srivastava et al.,2015]. A neural network is trained on a training input sample. A multilayered network has more than two layers. The layer between input layer and output layer is called as hidden layer and hence neurons in hidden layer are known as hidden neurons. Each layer is connected to each other and each connection has a weight associated with it. Weights and biases are initialized with random values [Mia et al.,2015]. The training sample of input is given to the input layer, which gives an output of the layer by applying transfer function and weight to the input value. The output is then presented as input to hidden layer having its own transfer function, weights and biases. In this way the output of the hidden layer,

called as the activation value of hidden neurons, is then presented to output layer which in turn gives the output of the network on the input sample. The calculated output is then compared with the target output. The difference between the two is calculated. The weights of the network are then changed by a small value, calculated by the training algorithm, to minimize the difference between the actual output and target output known as an error [Kamruzzaman & Hassan,2005].

The most minimized value of the error term is known as best performance value. After training and validation, network is trained to accept input samples and give the expected output value, i.e. value of selector attribute for that input sample.

A network is trained by a training algorithm which works on the principle of minimizing the error term. The error term is calculated at output layer by squaring the difference between the target output and the actual output. The goal of the training algorithm is to minimize this error term by changing representation of hidden units at each iteration, which makes it complex and needs more rules to explain. Researchers and scholars have worked on improvisation on training algorithms but not much on the internal representation of a network, the hidden layers and weights calculated at hidden neurons in comparison to training algorithms. Since in decompositional approach of rule extraction, rules are extracted from hidden units, therefore the number of rules mainly depends on hidden units and internal layer representation. During training hidden unit activation values can take their values anywhere in the space to achieve minimum squared error term calculated by the learning algorithm. Hence to get fewer rules, it is important to minimize the scattered activation values at hidden units in the network.

Rule extraction algorithm extracts rules from the trained network in terms of input and output [Kamruzzaman & Sarkar,2011]. It express the symbolic rules, for example, if for an input sample of a patient, (x1,x2,x3,x4,x5,x6,x7,x8,x9,x10) represents the values of 10 attributes of a patient then rules are expressed as:

If x1 (relational operator) v1 and x2 (relational operator) v2 and x3 (relational operator) v3 ……. and x10 (relational operator) v10 then

The patient will be a liver patient where vi is a numeric value for that attribute.

Since an input data sample can have many combinations of the values of attributes, the generalized number of rules should be minimum for taking decision on given set of values of attributes that whether the sample will be a patient or not, without affecting the accuracy of the decision.

The proposed rule extraction method, efforts to minimize the number of rules extracted from the network without affecting the classification accuracy.

2. The Distance Term

The three layered feed forward network is simulated and trained on the input samples.

The proposed method [Srivastava et al.,2015] follows the decompositional approach of rule extraction in which first rules are extracted between hidden units and output and then between input and hidden units. Combining the two gives rules in terms of input and output [Huynh & Reggia, 2011]. Since the number of rules mainly depends on hidden units and internal layer representation, the proposed method attempts to extract a fewer rule at hidden layer. For an input sample xp,

the activation value of a hidden unit Hi is represented as apHi and actual output as op and target output as tp. The activation values of hidden units will range from 0 to 1 after the logarithmic transfer function and target output will be either 0 or 1. For N hidden units, the activation values will be apH1 to apHN. The proposed method, calculates the Euclidean distances between hidden unit's activation value. For a given input, if the difference between two hidden unit's activation value is more, it means the two hidden units tends to give different output values for the same input. The activation value which is not clustered with other hidden unit's activation value will have a larger distance value from them in comparison to others (not clustered with others, but is in proximity), which represents that this hidden unit is tending to an intermediate output value. Such activation value is pushed towards the clustered hidden units to eliminate the rule with that intermediate hidden unit value. If the distance is more than 1, represents that the hidden unit's activation value is tending to a different output value in comparison to other clustered hidden unit activation values. Such scattered activation value is not pushed since it contributes to a different output, which should also be considered to maintain the accuracy of the rules. For example, if there are three hidden units, the activation values for an input sample is 0.245, 0.319 and 0.498. All three values are equidistant to each other and not significantly far to each other. Hence, these values will not be moved closer for accuracy of classification rules. If these values are 0.926, 0.899 and 0.513 then 0.513 will be moved to the calculated distance. In this way all three will contribute to only one rule. As explained earlier, the error gives the difference between the actual output and the target output for an input sample. If the best performance can be more minimized, the network output will be more closer to target output hence less scattered intermediate values. This will also contribute in reducing the number of rules extracted from the network. The proposed algorithm attempts to work on the same principle to contribute in extracting fewer rules from the network.

3. Data Sets

The proposed method is experimented on five standard datasets from UCI machine learning repository. ILPD This data set contains 416 liver patient records and 167 non liver patient records. Selector is a class label used to divide into groups(liver patient or not i.e 1 for patient and 0 for non). This data set contains 441 male patient records and 142 female patient records, total 583. A person will have some values for each of these 10 attributes WAVEFORM data set consists of 5000 instances of waves. Each wave is characterized by 21 continuous inputs with noise. The problem is to classify these waves into one of three classes. ARRHYTHMIA This database contains 279 attributes, 206 of which are linear valued and the rest are nominal. The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. The instances are divided randomly into three sets: 80 percent for training, 20 percent for testing, and 20 percent for validation. CTG (CARDIOTOCO-GRAPHIC DATA) The dataset consists of measurements of fetal heart rate

(FHR) and uterine contraction (UC) features on cardiotocograms classified by expert obstetricians. 2126 fetal cardiotocograms (CTGs) were automatically processed and the respective diagnostic features measured. The CTGs were also classified by three expert obstetricians and a consensus classification label assigned to each of them. Classification was both with respect to a morphologic pattern (A, B, C. ...) and to a fetal state (N, S, P). Therefore the dataset can be used either for 10-class or 3-class experiments. Here it is considered with 10 classes. The instances are divided randomly into three sets: 80 percent for training, 20 percent for testing, and 20 percent for validation. A three layer feed forward neural network with 11 hidden units is trained on the data. The output of the layer is clustered into 10 groups corresponding to the 10 classes

.

DATA SET	NO.OF. ATTRIBUTESS	NO. OF CLASSES	NO. OF INSTANCES
ILPD	10	2	583
WAVE FORM	21	3	5000
ARRYTHMIA	279	16	452
CTG	21	10	2128
IMAGE SEG	18	7	2310

.

Table 1: Data Sets Used for Evaluation

Image Segmentation: Image data described by high-level numeric valued attributes, 7 classes. The instances were drawn randomly from a database of 7 outdoor images. From each dataset, 80 percent of the data is used for the training the training,10 percent is used for testing and 10 percent for validation. A three layer feed forward neural network with 8 hidden units is trained on the data. The output of the layer is clustered into 7 groups corresponding to the 7 classes.

4. **Experiment and Result**

The goal of this evaluation is to compare the number of rules extracted from a trained network when Distance Term is included in the hidden layer (experimental condition) versus the number when Distance Term is not included (control condition). It shows that training with Distance Term produces better separated encoding at the hidden layer, and thus would improve the performance of existing rule extraction methods. The effectiveness of the rule extraction

method is evaluated on the data sets having more than 7 classes to classify selected arbitrarily from the UCI Machine Learning Repository These are large and difficult data sets with many attributes and classes which has not given better results on previous research work[Huynh & Reggia, 2011].

DATA SET	ACTUAL	NEW	%age
ILPD	28.54	15	47.4
WAVE FORM	69.04	39.8	42.4
ARRYTHMIA	32.54839	28.41139	12.7
CTG	165.9231	159.8225	5
IMAGE SEG	84.57	80.68548	5

Table II. Experiment Result of Rule Extraction

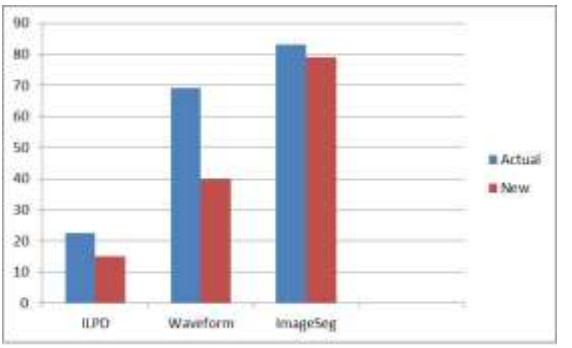

Figure 1

The hidden unit encodings learned by the neural network for five large dataset are used to illustrate proposed method. The proposed method was experimented on datasets for 150 iterations. The results are given below. The result shows that the number of rules extracted are significantly less in number and has not compromised on accuracy.

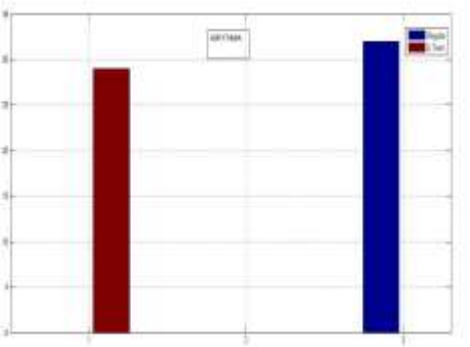

Figure 2

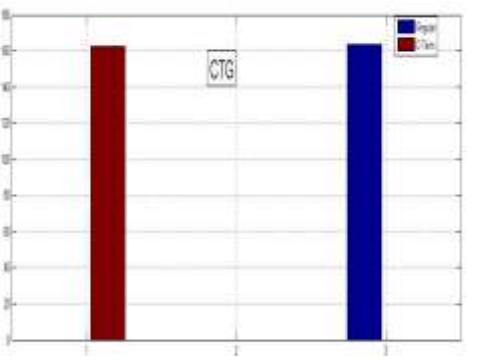

Figure 3

The experiments were repeated multiple times with different initial weights to rule out the effect of randomness and be more confident in the results. Accordingly, the results were averaged over 150 runs in which experiment runs started from the same initial conditions.

Table II shows details of all data sets and the actual number of rules and number of rules with distance term. The less number of rules is helpful to take a decision or conclude about the selector class for a given value of all attributes.

5. Observations and Analysis

Results are analyzed on many parameters namely:

- Number of Classes
  a. Distance moved.
  b. No of neurons moved.
- Number of Instances
- Number of Attributes
- Number of Iterations

S.No	Data Set	Actual	New	Avg Distance Shifted	Total Values Effected
1.	ILPD	28.54	15	1059	453
2.	WAVE FORM	69.04	39.8	49000	2367
3.	ARRYTHMIA	32.54839	28.41139	4100	1710
4.	CTG	163.9231	159.8225	96600	160
5.	IMAGE SEG	85.47	80.68548	45547.5	141

Table III: Experiment Results

Following Observations are made:

- ILPD Dataset has lowest number of attributes, classes and number of instances is also less. The result of ILPD is highest, best among all data sets.
- The waveform has more number of attributes in comparison to ILPD. Number of classes are just one more than ILPD but the number of Instances are almost 100 times more than ILPD and highest among all. This has effected on percentage of reduction in the number of rules.
- Arrythmia has the highest number of attribute and numbers of classes, but instances is lowest even less than ILPD. The percentage of reduction in the number of rules has reduced by more than one third.
- All three parameters of CTG and IMAGE SEGMENTATION are high in numbers. This fact has effected on percentage of reduced the number of rules drastically.
- Below table shows the average distance shifted and total number of neurons shifted towards their resulted class. The value of a neuron will fall in the interval of 0 to 1 for any of the classes of the dataset. Therefore, shifting of the value of a neuron towards its intended class has to be done very judiciously especially when the number of classes is more for a dataset. Waveform, CTG and Image Segmentation dataset have large average distance shifted in comparison to ILPD and Arrythmia but less number of neurons shifted.

6. Comparison of Results:

- ILPD and Waveform have the highest percentage of reduction in the number of rules, whereas CTG and Image Segmentation has the lowest. The difference between them is mainly the number of classes. All parameters of CTG and Image Segmentation are highest.
- Arrhythmia dataset has shown intermediate result. It has the largest number of attributes and classes, but less number of instances.
- ILPD and Waveform dataset has almost the same number of classes for classification. Waveform dataset has more number of attributes and instances in a comparison ILPD

dataset, resulted in slight fall in percentage of reduction in the number of rules.

- CTG and Image Segmentation dataset's all parameters are almost same and hence also has shown the same results.
- Since CTG and Image Segmentation dataset has a high number of attributes as well as instances, it resulted in a higher average distance shifted Since the number of classes is high, it resulted in fewer numbers of shifting neurons.

7. Conclusion

In this paper, five large dataset are used to experiment the performance of the enhanced rule extraction algorithm proposed by [Srivastava *et al.*, 2015]. The results are analyzed to conclude the effect of considered parameters on number of extracted rules. We can summarize the analysis by stating that above results show that the number of attributes and number of instances affects the reduction in the number of rules. The Algorithm works well on a dataset having a lower number of instances and attributes. Similarly classification into the least number of classes gives better results. The number of neurons shifted from one class to another also depends on the number of attributes and instances of the dataset. The number of classes does not affect significantly on it. Hence we can conclude that the reduction in rules depends on the number of values affected of neurons, which are again affected by the number of instances and attributes of the dataset. This analyses may contribute in the application of neural network classification.

8. References Section

A,Gupta. (1999).Generalized Analytic Rule Extraction for Feed forward Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, 11(6), 985-991.

S.M. Kamruzzaman and A. R. Hasan(2005). Rule Extraction using Artificial Neural Networks. ICTM.

S.M. Kamruzzaman and A.M. Jehad Sarkar(2011) . A new data mining scheme using artificial neural networks Sensors.11, 4622-4647.

T. Q. Huynh and J. A. Reggia (2011). Guiding Hidden Layer Representations for Improved Rule Extraction from Neural Networks, *IEEE, Neuralnetworks*, 22(2), 264-275.

V. Srivastava, C. Dhawale and S. Misra(2015). Enhanced Rule Extraction by Augmenting Internal Representation of Feed Forward Neural Network, *IEEE 978-1-4673-9354-6/15/$31.00 ©2015*

M. M. Mia, S. K. Biswas and M. C. Urmi(2015). An Algorithm For Training Multilayer Perceptron (MLP) For Image Reconstruction Using Neural Network Without Overfitting, IJSTR, 271-275. A. Asuncion and D.J. Newman, UCI Machine Learning Repository, http:==www.ics.uci.edu/mlearn=MLRepository:html:

L. Ai-sheng, Z.Qi. Automatic modulation classification based on the combination of clustering and neural network, Science Direct, 2011.