Object oriented approach to combined learning of decision tree and ADF GP

Ayahiko Niimi and Eiichiro Tazaki

Department of Control and Systems Engineering,
Toin University of Yokohama
1614 Kurogane-cho, Aoba-ku, Yokohama 225-8502, JAPAN
E-mail: tazaki@intlab.toin.ac.jp

Abstract

There are many learning methods for classification systems. Genetic programming (one of the methods) can change trees dynamically, but its learning speed is slow. Decision tree methods using C4.5 construct trees quickly, but the network may not classify correctly when the training data contains noise. For such problems, we proposed an object oriented approach, and a learning method that combines decision tree making method (C4.5) and genetic programming. To verify the validity of the proposed method, we developed two different medical diagnostic systems. One is a medical diagnostic system for the occurrence of hypertension, the other is for the meningoencephalitis. We compared the results of proposed method with prior ones.

1 Introduction

Various techniques are proposed for the construction of the inference system using classification learning. In general, the learning speed of a system using a genetic programming is slow. Moreover, both problem background knowledge and design skill are demanded of the system designer. However, a learning system which can acquire higher-order knowledge by adjusting to the environment can be constructed, because the structure is treated at the same time.

On the other hand, a learning system which uses the decision tree can be trained in a short time compared to other techniques. An effective network structure is constructed by the classification model is obtained by decision tree. But, there is a problem with deteriorated classification accuracy when the training data contains noise.

Each technique has advantages and disadvantages like this. In this paper, we propose an object oriented method to construct a classification system trained by combining various learning methods treated as objects. It is expected that learning will occur while mutually making up for the advantage and the disadvantage of each technique.

To verify the validity of the effectiveness of the proposed learning method, we test the two learning methods: the decision tree construction method (C4.5), and genetic programming with automatic function definition (ADF). This is applied to two different medical diagnostic systems. One is a medical diagnostic system for the occurrence of hypertension, the other is for meningoencephalitis. We compared the results of the proposed method with prior approaches using only one learning method.

The effectiveness of combining techniques has been verified by the study of neural network training using decision tree construction method and back-propagation learning method[1], and by the study of neural network training using decision tree construction method and genetic programming[2].

2 Decision tree construction algorithm

There is a method of inductively constructing a model by examining the recorded classification data and generalizing a specific example. The classification learning by decision tree can achieve a certain classification ability in a comparatively short time. C4.5 is one of decision tree construction methods, and it classifies data based on the gain criterion which selects a test to maximize expected information gain. As a result, important attributes can be collected at the root of the decision tree. Moreover, the algorithm contains branch pruning by estimating error rates to prevent the construction of excessively classified decision trees.

2.1 The algorithm of C4.5

The decision tree construction with C4.5 follows the following procedures[3].

1. Construction of initial decision tree.
2. Branch pruning of constructed decision tree.

2.2 Decision tree construction and number of training data

In the decision tree construction method by C4.5, the decision tree is constructed with the recurrent division of the training data. Therefore, there is a possibility of constructing a different decision tree when the number of training data is modified. In general, using a large number of training data tends to construct a more complex decision tree.

3 Genetic programming algorithm

Genetic Programming(GP) is a learning method based on the theory of natural evolution, and the flow of the algorithm is similar to that of Genetic Algorithm(GA). The difference with GA is that in GP the chromosome expression has been extended to express structure using function nodes and terminal nodes. In this paper, the tree structure was used to express the decision tree.

3.1 The algorithm of GP

The decision tree construction with GP follows the following procedures.

1. An initial population is generated from a random grammar of the function nodes and the terminal nodes defined for each problem domain.
2. The fitness value, which relates to the problem solving ability, for each individual of the GP population is calculated.
3. The next generation is generated by genetic operations.
   (a) The individual is copied by fitness value (reproduction).
   (b) A new individual is generated by intersection (crossover).
   (c) A new individual is generated by random change (mutation).
4. If the termination condition is met, then the process exits. Otherwise, the process repeats from the calculation of fitness value in step 2.

3.2 Automatic function definition

Ordinarily, there is no method of adequately controlling the growth of the tree, because GP does not evaluate the size of the tree. Therefore, during the course of the search the tree may become overly deep and complex, or may settle to a too simple tree . There has been research on methods to have the program define functions itself for efficient use. One of the approaches is automatic function definition (or Automatically Defined Function:ADF), and this is achieved by adding the gene expression for the function definition to normal GP[4]. By implementing ADF, a more compact program can be produced, and the number of generation cycles can be reduced. More than one ADF can be defined in one individual.

3.3 Fitness function

In addition, the growth of the tree is controlled by evaluating the size of the tree. For example, an approach based on minimum description length (MDL) has been proposed concerning the evaluation of the size of the tree. In this paper, we used the classification success ratio and the number of composed nodes, to define the fitness of the individual.

$$f_{fitness(n)} = \alpha f_{hits(n)} + (1 - \alpha)\, f_{nodes(n)}$$

α is weight defined by $(0 \leqq \alpha \leqq 1)$.

Here, two things are expected. One is that accuracy of the decision tree is raised while avoiding over-training in GP. And the other is that the decision tree generated can be made comparatively compact. The inference accuracy and rule size of the best individual is influenced by the weighting of success rate or node size.

4 Object oriented training approach

In an object oriented design, a program is built from independent information processing units, called objects, and information is processed by exchanging messages between objects. The object receives input information, performs a series of processes, and sends output information. In this case, the object's own internal information is hidden. Therefore, the external data flow can be designed independently from the internal processing.

4.1 Learning by combination of decision tree and GP

The classification learning by the decision tree can be trained in a short time, but when the noise is contained in the training data, the classification ability is rapidly deteriorated. On the other hand, a high classification ability is obtained by GP through training structural information, but more learning time is needed as the degree of learning freedom increases.

For such problems, we propose an object oriented approach to the learning method that combines the decision tree method (C4.5) and genetic programming. The proposed method consists of the following three steps:

1. First, using C4.5, construct appropriate decision trees.
2. Next, generate the genetic programming population which includes initial individuals converted from the decision trees.
3. Train the genetic program to construct the classification system.

The proposed method simplifies the construction of classification systems. It is observed that the discursive accuracy of classification becomes highly improved by the emergent property of interaction between two combined methods.

4.1.1 Initialization of GP using decision trees

When the decision tree is taken into the initial population of GP, it is necessary that variety in the initial population is not upheld[5]. For that purpose, we created decision trees by C4.5 using different number of training data, to ensure a variety of patterns taken into the initial population.

5 Applications to medical diagnostic system

To verify the validity of the proposed method, we developed two different medical diagnostic systems. One is a medical diagnostic system for the occurrence of hypertension, the other is for meningoencephalitis. We compared the results of the proposed method with prior ones.

5.1 Medical diagnostic system for the occurrence of hypertension

The database used for the occurrence of hypertension contains fifteen input terms and one output term. There are two kinds of Intermediate assumptions between the input terms and the output term[6]. Among the input terms, ten terms are categorized into a biochemical test related to the measurement of blood pressure for past five years, and the rest are "Sex", "Age", "Obesity Index", "γ-GTP", and "Volume of Consuming Alcohol". One output term represents whether the patient has had an attack of hypertension for the input record. The database has 1024 patient records. In this paper, we selected 100 occurrence data and 100 non-occurence data by random sampling, and this was used as the training data.

The parameters for GP used the following. (Refer to Table 1.) In this experiment, only a small percentage of GP individuals could be used as decision trees due to the large number of node types and large freedom for tree construction. Moreover, most of the input data have continuous value attributes. This seems to have caused the decrease in the inference accuracy of the GP search close to that of a random search. It was confirmed that the decision tree taken in as an initial individual was succeeded to the best individual, and it can be concluded that this influenced the improvement of the learning efficiency. (The results shown in Table 2.)

Table 1: Parameters of GP

GP population	500
Reproduction probability	0.1
Intersection probability	0.8
Mutation probability	0.1
Selection method	Tournament method
Function node	IFLTH, IFEQ, $*$, $/$, $+$, $-$, ADF0, ADF1
Terminal node	Attribute value of database such as SEX and AGE (15 attributes), P, N, R
Number of training data	P(100), N(100)
Number of test data	1024

IFLTH,IFEQ: if less than $(<)$,if equal to$(=)$
ADF0, ADF1: the function definition gene expanded by ADF
R: randomly generated constant
P, N: occurence (P), no-occurence (N)

Table 2: experiment result (reasoning precision) by each technique.

	training data (%)	all data (%)	nodes (%)	generations
C4.5	98.5	77.1	29	—
ADF-GP	75.0	66.9	148	14328
C4.5 +ADF-GP	96.0	81.4	17	41

5.2 Medical diagnostic system for meningoencephalitis

The meningoencephalitis database is a medical treatment database concerning the discrimination diagnosis of meningoencephalitis. It consists of 140 patients. The database is described by 34 attribute about the past illness history, physical examinations, laboratory examinations, diagnosis, therapies, clinical courses, final status, and risk factors. The two classifications were bacillus and virus meningitis[7]. In this paper, 32 attributes regarding sex, ages, etc. were used as well.

The parameters for GP used the following. (Refer to Table 3.) For this experiment, all approaches achieved high accuracy, the medical database had removed noise well, and many attributes have discrete values. For the construction of decision tree by ADF-GP only, the inference accuracy did not increase quickly, but for construction of decision tree by combined ADF-GP and C4.5, the decision tree reached high accuracy comparatively quickly. It was confirmed that the decision tree taken in as an initial individual was succeeded to to the best individual, and influenced the improvement of the learning efficiency. (The results shown in Table 4.)

6 Conclusions

In this paper, we proposed an inference system construction method based on an object oriented approach, combining two or more learning methods. We compared the results of three methods (using C4.5 only, using ADF-GP only, and using combined C4.5 and ADF-GP using the proposed method).

As a result, an improvement of learning efficiency was seen. It can be concluded that the proposed technique is an effective method for the improvement of learning efficiency of an inference system.

Table 3: Parameters of GP

GP population	500
Reproduction probability	0.1
Intersection probability	0.8
Mutation probability	0.1
Selection method	Tournament method
Function node	IFLTH, IFEQ, $*$, $/$, $+$, $-$, ADF0, ADF1
Terminal node	Attribute value of database such as SEX and AGE (32 attributes), VIRUS, BACTERIA, R
Number of training data	VIRUS(49), BACTERIA(21)
Number of test data	140

IFLTH,IFEQ: if less than $(<)$,if equal to$(=)$
ADF0, ADF1: the function definition gene expanded by ADF
R: randomly generated constant
VIRUS,BACTERIA: Virus meningitis and bacillus meningitis

Table 4: Comparison of generated decision trees (number of nodes and inference accuracy) by each technique

	training data (%)	all data (%)	nodes (%)	generations
C4.5	98.6	94.3	11	—
ADF-GP	88.6	82.9	21	7673
C4.5 +ADF-GP	95.7	97.1	11	686

Acknowledgements

The authors would like to express their thanks to Dr. Katusmi Yoshida at the Department of Preventive Medicine, St. Marianna University School of Medicine for the medical dataset for the occurrence of hypertension, and helpful discussions.

The authors would like to express their thanks to Dr. Shusaku Tsumoto at the Medical Research Institute, Tokyo Medical and Dental University for the meningoencephalitis medical dataset and helpful discussions.

References

[1] A. Banerjee, R. Greiner(ed.), et.al., Initializing Neural Networks Using Decision Trees, Computational Learning Theory and Natural Learning Systems, pp.3–15, 1997

[2] N.Matsumoto, E.Tazaki, Emergence of Learning Rule in Neural Networks Using Genetic Programming Combined with Decision Tree, Proceeding of IEEE International Conference on Systems, Man, and Cybernetics (SMC'98), pp.1801–1805, 1998

[3] J. R. Quinlan, C4.5:Programs for Machine Learning,Morgan Kaufmann Publishers,1995

[4] J. R. Koza, K. E. Kinner(ed.), et.al, Scalable Learning in Genetic Programming Using Automatic Function Definition, Advances in Genetic Programming, pp.99–117, 1994

[5] A. Niimi, E. Tazaki, Combined Learning Method of Decision Tree and ADF GP by Object Oriented Approach (In Japanese), Proceedings of 42nd KBS meeting,pp.75–82, Japan AI Society,1999

[6] T. Ichimura, K. Ooba, et.al., Knowledge Based Approach to Structure Level Adaptation of Neural Networks, Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC'97),pp548–553,1997

[7] S. Tsumoto, et. al., Comparison of Data Mining Methods using Common Medical Datasets, Proceeding of Data Mining and Knowledge Discovery in Data Science(ISM Symposium),pp.63–72,1999