Rule Discovery Technique Using Genetic Programming Combined with Apriori Algorithm

Ayahiko Niimi[1] and Eiichiro Tazaki[1]

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

1 Introduction

Various techniques have been proposed for rule discovery using classification learning. In general, the learning speed of a system using genetic programming (GP) [1] is slow. 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, there is the Apriori algorithm [2], a rule generating technique for large databases. This is an association rule algorithm. The Apriori algorithm uses two values for rule construction: a support value and a confidence value. Depending on the setting of each index threshold, the search space can be reduced, or the candidate number of association rules can be increased. However, experience is necessary for setting an effective threshold.

Both techniques have advantages and disadvantages as above. In this paper, we propose a rule discovery technique for databases using genetic programming combined with the Apriori algorithm. By using the combined rule generation learning method, it is expected to construct a system which can search for flexible rules in large databases.

2 Proposed Rule Discovery Technique

We propose a rule discovery technique which combines GP with the Apriori algorithm. By combining each technique, it is expected to increase the efficiency of the search for flexible rules in large databases.

The following steps are proposed for the rule discovery technique.

1. First, the Apriori algorithm generates the association rule.
2. Next, the generated association rules are converted into decision trees which are taken in as initial individuals of GP. The decision trees are trained by GP learning.
3. The final decision tree is converted into classification rules.

This allows effective schema to be contained in the initial individuals of GP. As a result, it is expected to improve the GP's learning speed and its classification accuracy.

S. Arikawa and S. Morishita (Eds.): DS 2000, LNAI 1967, pp. 273-278, 2000.

For conversion from the association rule into decision trees, we use the following procedures.

1. For the first process, the route of the decision tree is constructed, assuming the conditions of the association rule as the attribute-based tests of the decision tree.
2. In the next process, the conclusions of the association rule is appended on the terminal node of this route.
3. Finally, the terminal nodes which are not defined by the association rule are assigned candidate nodes at random.

For conversion from the GP's decision tree to the classification rule, we use the process proposed by Quinlan [3].

3 Experiments

To verify the validity of the proposed method, we applied it to the house-votes data from UCI Machine Learning Repository [4], and medical database for occurrence of hypertension [5]. From here on all occurrence of GP uses Automatically Defined Function Genetic Programming (ADF-GP) [1] including the proposed method. In the proposed method, we took the association rule generated by Apriori algorithm as initial individuals of GP. We compared the results of the proposed method against GP. We use house-votes database as small test database expressed by discrete values, and hypertension database as large test expressed by continuous values.

3.1 Application to house-votes Data

For evaluation, we used house-votes data from UCI Machine Learning Repository [4]. We compared the results of the proposed method with GP. The evaluation data contains 16 attributes and 2 classes. The attributes are for example "handicapped-infants" and "water-project-cost-sharing " etc. They are expressed by 3 values: "y ", "n", and "?". And the 2 classes are "democrat" and "republican ". 50 cases out of the total 435 data of house-votes were used for training data.

We extracted the association rule from the database by the Apriori algorithm. We applied the Apriori algorithm to a data set excluding data with the "?" value, because "?" value means "others". In the following experiment, we used minimum support value (= 30) and minimum confidence value (= 90). As a result of the experiment, 75 rules were generated.

Next, the above generated 75 rules were taken into the initial individual. The result of the evaluation of each fitness value is shown in figure 1, and the result of best individual is shown in table 1.

By using GP, inference accuracy did not improve rapidly. However, the proposed method showed fast learning and achieved high accuracy. Comparing the

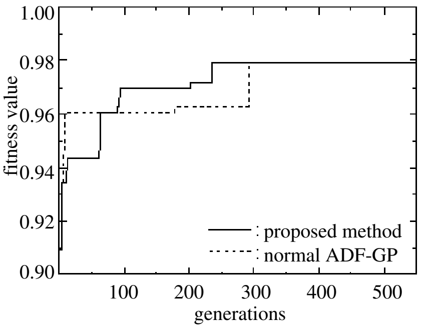

Fig. 1. Evaluation of Each Fitness Value

Table 1. Experiment Best Individuals Result (House-votes).

	training (%)	all data (%)	nodes	depths	generations
ADF-GP	100.0	86.0	11	3	586
Apriori + ADF-GP	98.0	92.9	9	2	235

best individual results, the proposed method showed better results than GP, except for accuracy against the training data. Concerning the results of training data, GP may have shown overfitting, but its proof could not be obtained by only this result.

The rules were converted from the constructed decision tree removing invalid rules and meaningless rules. The rules' total accuracy was 94.8%.

3.2 Application to Medical Database

We applied a medical diagnostic system for the occurrence of hypertension. We compared the results of proposed method with GP. Most of the data values are expressed as continuous values, and the size of the database is larger than the house-votes database.

The occurrence of hypertension database contains 15 input terms and 1 output term. There are 2 kinds of intermediate assumptions between the input terms and the output term[6]. Among the input terms, 10 terms are categorized into a biochemical test related to the measurement of blood pressure for past five years, and the other terms are "Sex", "Age", "Obesity Index", "γ-GTP", and "Volume of Alcohol Consumption". 1 output term represents whether the patient has an attack of hypertension for the input record. The database has

1024 patient records. In this paper, we selected 100 occurrence data and 100 no-occurrence data by random sampling, and this was used as the training data.

The association rule has been extracted from the database by the Apriori algorithm. The Apriori algorithm was used after these attributes had been converted into binary attributes using the average of each data, because the continuous value attributes were included in this database. To search for the relationship between the minimum support value and the minimum confidence value and the number of rules, we experimented with the threshold patterns. (Refer to the result table 2) In the following experiment, we used minimum support value (= 30) and minimum confidence value (= 90).

Table 2. Relations between Thresholds and Number of Rules

Minimum Support Value	Minimum Confidence Value	Rules
25	75	396
30	75	125
25	90	187
30	90	33

Next, the 33 rules generated by the Apriori algorithm were taken into the initial individual. The result of best individual is shown in table 3.

By using GP, inference accuracy did not improve rapidly. However, the proposed method showed fast learning and achieved high accuracy.

Table 3. Experiment Best Individuals Result (Hypertension).

	training (%)	all data (%)	nodes	depths	generations
ADF-GP	89.5	66.3	41	6	18553
Apriori + ADF-GP	89.5	74.9	49	4	671

When the rules were converted from the decision tree, invalid rules and meaningless rules were removed. Each ratio of the number of effective rules to generation rules was 37.5% (by GP) and 50.0% (by proposed method). (Table 4 shows 3 rules generated with each technique, chosen by the highest support value.)

By using GP, many invalid rules and many rules which were difficult to interpret were generated. Compared to GP, the proposed method showed decrease in the support value and improvement in accuracy. The proposed technique improved the ratio of effective rules and the accuracy.

Table 4. Comparison of Generated Rules (Size and Inference Accuracy)

Technique	Size	Support Value(%)	Wrong(%)
ADF-GP	4	41.4	48.4
	2	36.0	24.1
	4	22.6	8.2
Apriori+ADF-GP	2	17.7	39.2
	4	15.5	13.2
	4	13.8	19.2

4 Concluding Remarks

In this paper, we proposed a rule discovery technique for databases using genetic programming combined with association rule algorithms.

In the future, we will study the following 4 topics related to the proposed method. The first topic is to apply the method to other verifications. The second topic is to further research the conversion algorithm from the association rule to a decision tree with high accuracy. The third topic is to extend the proposed method to multi-value classification problems. The fourth topic is to do a theoretical analysis about the mechanism of the overfitting.

References

1. J. R. Koza, Genetic Programming II, MIT Press, 1994.
2. M. Terabe, O. Katai, T. Sawaragi, T. Washio, H. Motoda. Attribute Generation Based on Association Rules. Journal of Japanese Society for Artificial Intelligence, Vol.15 No.1,pp.187–197, 2000. (In Japanese)
3. J. R. Quinlan: C4.5: Programs for Machine Learning, Morgan Kaufman Publishers, 1993.
4. C. L. Blake, C. J. Merz. UCI Repository of machine learning databases. [http://www.ics.uci.edu/˜mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science, 1998.
5. A. Niimi, E. Tazaki: Object oriented approach to combined learning of decision tree and ADF GP, 1999 International Joint Conference on Neural Networks, Washington DC, USA, 1999.