Data Mining using Genetic Network Programming

Takeshi FUKUDA* Kaoru SHIMADA* Kotaro HIRASAWA* Takayuki FURUZUKI*
*Graduate School of Information, Production and Systems, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka 808-0135, Japan
e-mail: kurt2@moegi.waseda.jp

Abstract

Recently, Data Management System (DMS) is used in various fields, and the importance of data mining extracting important rules from a large-scale database has been recognized widely. Although the conventional data mining using statistical techniques has been developed, a novel algorithm of data mining using Genetic Network Programming (GNP)is proposed in this paper. The proposed system using GNP with directed graph structures can extract automatically the correlation rules showing the importance of data. GNP has been proposed and studied as a new method of evolutionary computations. GNP is constructed by the network structure whose gene consists of the directed graph, so it is possible to search solutions effectively by the implicit memory function of the network structure. Until now, the applicability and availability of GNP to the real-world applications have not been studied, whereas it has been applied to virtual-world examples such as Tile-world and its effectiveness has been proved through the comparison with GP. We describe a method to extract correlation rules automatically using support and confidence, and experimental results are described to show the effectiveness of the proposal method.

Keywords: Data Mining, Correlation Rule, Genetic Network Programming, Evolutional Algorithm.

1 Introduction

Information is spreading in the information-oriented society almost every day. In this circumstance, Data Management System (DMS) has been attracted recently. People interested in DMS have not only managed the data but tried to extract knowledge from the data, and this has become important theme in these days. In this sense, data mining capable of extracting hidden knowledge from an enormous database has been noticed recently. Database techniques using statistical techniques such as decision tree, correlation rule, and cluster analysis have been proposed in order to carry out data mining efficiently.[1][5]

Finding the correlation rules[1][5], that is, discovering the relevance or pattern existing in the set of items of the records, is one of the methods which can be easily understood.

But, the cost of extracting the correlation rules using the traditional statistical techniques is fairly high.

In this paper, we propose a novel data mining technique using Genetic Network Programming (GNP)[2] whose gene has directed graph structures in order to overcome the disadvantages of the conventional methods.

A complex mathematical algorithm could not solve the problems caused by the large size of database but an evolutional algorithm could do it because GNP has such features as *1)* reusing the nodes and *2)* having the directed graph structure. In other words, GNP can extract correlation rules efficiently using the directed graph structure and keeping the size of the individuals fixed due to the reused nodes.

The fitness of GNP having a number of correlation rules is calculated and GNP is evolved by the calculated fitness.

Data mining and GNP are briefly discussed in section 2 and 3, respectively. The basic structure of the proposed method is described in section 4, followed by simulation results in section 5, and the conclusions in section 6.

2 Data Mining (Market Basket Analysis)

In this section, the market basket analysis is reviewed briefly, which is used in the simulation of this paper as the benchmarking problem.

When we deal with the shopping cart filled with goods in a supermarket, we can talk about the market basket analysis.[1][5] The shopping cart tells what a

customer buys. Each customer buys a quality of different items because the characteristics of customers are different from each other. Therefore, we can get a lot of information through them.

We can analyze the purchase lists of buyers in the market basket analysis to understand the general trend of the purchase of customers. The market basket analysis can give us the insight about the information on buying goods and can suggest us the new design of the shopping site because it shows us what combination of the goods is sold well. Furthermore it can show us where better sites for special goods are.

Merits of the market basket analysis are the availability and clearness of the results expressed by correlation rules. The correlation rules have intuitional properties because they express what kinds of relations the goods on the shelves have each other. Namely, the merits of finding the correlation rules by the conventional methods are:

1. It is easy to understand the correlation results clearly.
2. It is powerful to analyze the complicated data.
3. It is suitable to the data with variable size.
4. It is simpler than other algorithms for analyzing the data.

However they have several shortcomings:

1. The cost of calculating correlation rules increases exponentially with the size of problems.
2. It is difficult to deal with rare items.

We will show that the proposal method can overcome the above shortcomings by analyzing the market basket analysis using GNP in this paper.

3 Genetic Network Programming

In this section, Genetic Network Programming (GNP) is explained in detail. Basically, GNP is an extension of GP in terms of gene structures. The original idea is based on the more general representation ability of graphs than that of trees.[2]

3.1 Basic structure of GNP

The basic structure of GNP is shown in Fig.1. As shown in Fig.1, the directed graph structure is used to represent individuals. GNP is composed of plural

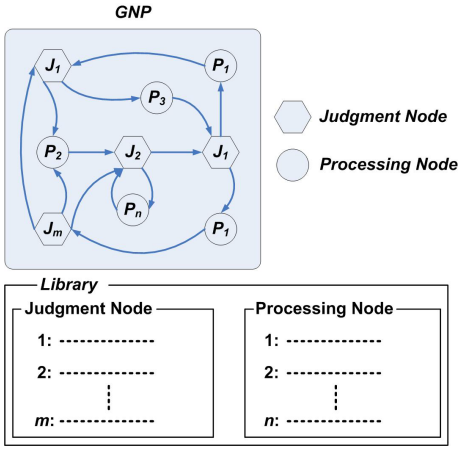

Figure 1: The basic structure fo GNP individual

nodes which are roughly classified into two kinds of nodes: `JUDGEMENT NODE` and `PROCESSING NODE`.

`JUDGEMENT NODE`s correspond nearly to elementary functions of GP and `PROCESSING NODE`s correspond almost to terminal symbols of GP. `JUDGEMENT NODE`s are the set of $J_1, J_2, \cdots, J_m$, which work as some kinds of judging functions. On the other hand, `PROCESSING NODE`s are denoted by the set of $P_1, P_2, \cdots, P_n$, which work as some kinds of action/processing functions. The practical roles of these nodes are predefined and stored in the library by supervisors.[3]

GP's elementary functions and terminal symbols are repeatedly used in a tree structure. In the same way, there are some $J_{1s}, J_{2s}, P_{1s}, P_{2s}$ and so on in GNP as shown in Fig.1. These `JUDGEMENT NODE`s and `PROCESSING NODE`s are the essential elements of GNP. The number of these nodes may be determined as a result of evolution like GP. Actually, GNP can use this strategy, in other words, GNP can adopt evolving the genotypes with variable number of nodes, but in this paper GNP evolves only the networks with the predefined number of nodes. It would be better to say that GNP here evolves the genotypes with fixed number of nodes. We set the number of each node in GNP equal to each other, e.g., $J_1 \times 3, J_2 \times 3, \cdots, P_1 \times 3, P_2 \times 3, \cdots$, and so on.

Additional specific nodes, start node S, is involved in GNP. Start node indicates the start point of GNP, which corresponds to GP's root node.

Once GNP is booted up, the execution starts from the start node, then the next node to be executed is determined according to the connection from the current activated node. If the activated node is `JUDGEMENT`

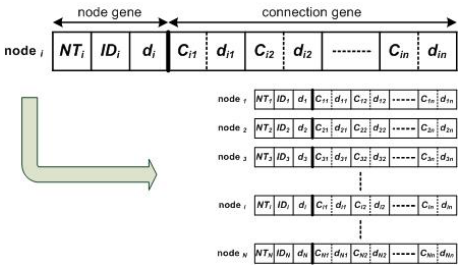

Figure 2: The genotype expression of GNP node

NODE, the next node is determined by the judgement at the activated JUDGEMENT NODE. When PROCESSING NODE is executed, the next node is uniquely determined by the single connection from PROCESSING NODEs.

The genotype expression of GNP node[4] is shown in Fig.2. This describes the gene of node i, then the set of these genes represents the genotype of GNP individuals. All variables in these genes are described by integer. NT_i describes the node type, $NT_i = 0$ when the node i is JUDGEMENT NODE, $NT_i = 1$ when the node i is PROCESSING NODE. ID_i is an identification number, e.g., $NT_i = 0$ and $ID_i = 1$ mean node i is J_1. $C_{i1}, C_{i2}, \cdots$, denote the nodes which are connected from node i firstly, secondly, $\cdots$, and so on depending on the arguments of node i. The total number of connection genes depends on the arity of the node's function. d_i and d_{ij} are the delay time. They are the time required to execute the processing of node i and delay time from node i to node C_{ij}, respectively. GNP can become materialized more realistically by setting these delays.

3.2 Genetic operator of GNP

The following genetic operators[4] shown in Fig.3 and 4 are used in GNP. Mutation operator affects one individual. All the connections of each node are changed randomly by mutation rate of P_m (shadowed in Fig.3). Crossover operator affects two parent individuals All the connections of the uniformly selected corresponding nodes in two parents are swapped each other by crossover rate of P_c between the two parents (shadowed in Fig.4). GNP evolves the fixed number of nodes, as I mentioned before, crossover is applied to the corresponding nodes selected uniformly in two parent genotypes as shown in Fig.4.[2][3][4]

Note that these genetic operators will not change any node functions, they only change the connection among the nodes. Therefore GNP doesn't evolve the functions of the nodes, but evolves the connections between nodes.

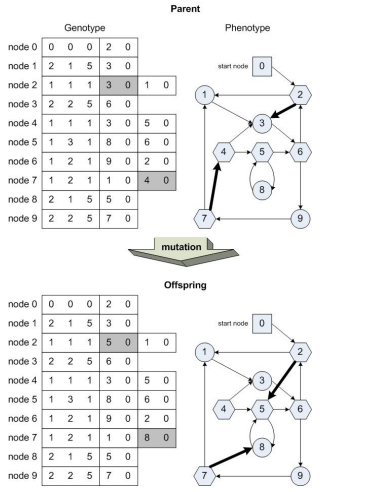

Figure 3: Mutation of GNP

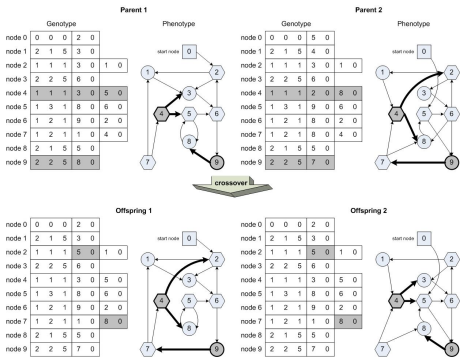

Figure 4: Crossover in GNP

4 Data Mining Structure using GNP

In this section, the proposed method for extracting correlation rules is described. First, we describe the fundamental concept dealing with the application of GNP to data mining.

1. Set the structural condition of GNP to extract correlation rules effectively.

2. Express a correlation rule by transition from the group of judgment nodes to the group of processing nodes in GNP.

3. Determine if the extracted rules are good enough or not using the support and the confidence.

4.1 Expression of Correlation Rules with GNP

Here, how to express the correlation rules using GNP is stated. The correlation rule shows the following relation between attributes in database,

$$if\ X\ then\ Y\ (X \subset I,\ Y \subset I,\ I : item\ set) \quad (1)$$

`JUDGEMENT NODE` corresponds to each element of set X, while `PROCESSING NODE` deal with each element of set Y, when a correlation rule is expressed by GNP. We define a correlation rule as the transition from the starting node S to the `PROCESSING NODE` following a `JUDGEMENT NODE` J. For example, a correlation rule of $if\ A, B\ then\ C, D$ is expressed by the node connection like Fig.5.

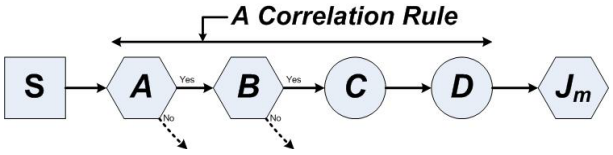

Figure 5: The expression of the correlation rule using nodes in GNP

4.2 Support and Confidence

When the correlation rule is expressed by $(X \Rightarrow Y)$, where X is called antecedent and Y is called consequent, we define $Support$ as the ratio of the records satisfying both X and Y, and $Confidence$ as the ratio of records satisfying Y among the records containing X.[1][5] The following $Support$ and $Confidence$ are used as the index to evaluate the importance of the correlation rule.

$$\begin{aligned} Support(X \Rightarrow Y) \\ Confidence(X \Rightarrow Y) \end{aligned} \quad (2)$$

For example, $Support$ of $(X(A, C) \Rightarrow Y(D))$ is 0.25, and the $Confidence$ is 0.5 in Table.1. In the case of $(X(B, C) \Rightarrow Y(E))$, $Support$ is 0.5 and $Confidence$ of it is 1.0.

Comparing with the conventional data mining method (apriori algorithm[1]) that extracts all the rules having more than minimum $Support$ and $Confidence$ values in database, GNP has a feature that it can extract correlation rules having high fitness considering the structural condition of GNP.

Table 1: Database example

Database					
TID	ITEM				
	A	B	C	D	E
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1

4.3 Fitness

Individuals of GNP are evolved by an evolutionary algorithm using $Support$ and $Confidence$ as the index of fitness. The fitness of the proposed algorithm uses the following:

$$Fitness = \frac{1}{|L|}\left(\sum_{l \in L} Support(l) \times Confidence(l)\right) \quad (3)$$

L	:	set of the number of rules in GNP
Support(l)	:	Support of rule l in GNP
Confidence(l)	:	Confidence of rule l in GNP

4.4 Extracting correlation rules by GNP

We set up several constraints to extract the correlation rules effectively in designing GNP. First, GNP reuses the nodes. If the node already used is reused at `JUDGEMENT NODE`s, a loop is created among `JUDGEMENT NODE`s. To overcome the loop problem, we set up several `JUDGEMENT NODE`s in GNP which express the same item.

In addition, we set several starting nodes connecting to `JUDGEMENT NODE`s, which means that all the rules are created by searching all the tuples in the database from the starting nodes.

The method to generate correlation rules using GNP is as follows:

1. Generate if part of the rule by searching the tuples in the database.

2. Generate then part of the rule by transfering the `PROCESSING NODE`s in GNP.

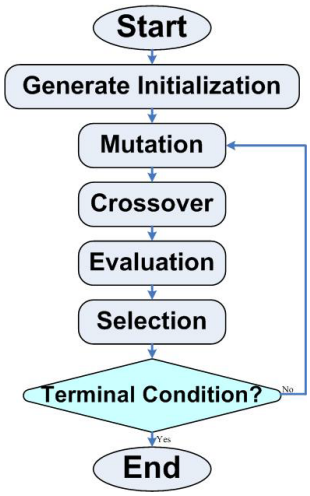

Figure 6: The flow chart of the proposed algorithm

We use a general evolutionary algorithm and select the better individuals for the next generation.

The flow chart of the proposed algorithm is as follow (See Fig.6).

1. Initialize GNP randomly.
2. Apply genetic operations to GNP.
3. Generate correlation rules using GNP.
4. Evaluate the generated rules by *Support* and $Confidence$ of the rules.
5. Calculate the fitness of each individual.
6. Select the better individuals to the next generation.
7. Move to 2, until a terminal condition meets.

5 Simulation

The effectiveness of the proposed algorithm is studied by extracting the correlation rules from the database having 10 attributes.

5.1 Simulation Parameters

200 individuals of GNP are used considering that many rules should be extracted in a short running time. The mutation probability and the crossover probability were set at 0.01 and 0.1 respectively[2][3][4]. We also used the elite strategy to move the elite individual to the next generation.[2]

5.2 Simulation Results

Database is made up of 10 attributes (from A to J) and 100 records. Each attribute is a binary code. Assuming the market basket analysis as an benchmarking problem, 0 denotes that the customer doesn't buy the item and 1 indicates that he buys the item. Moreover, the database used in the simulation is made artificially. For example, we set five artificial rules in the database like the records from 1 to 6 are the ones for extract-

Table 2: *Support* and $Confidence$ of the given 5 kinds of rules

	Correlation Rule	Support	Confidence
1	$A \Rightarrow J$	0.050000	0.500000
2	$B \Rightarrow E$	0.060000	0.461538
3	$C \Rightarrow D$	0.070000	0.636364
4	$F \Rightarrow G$	0.060000	0.666667
5	$H \Rightarrow I$	0.060000	0.428571

Table 3: The results wiht evolutionary algorithm

	Correlation Rule	Support	Confidence
1	$A \Rightarrow J$	0.050000	0.500000
2	$B \Rightarrow E$	0.060000	0.461538
3	$C \Rightarrow D$	0.070000	0.636364
4	$F \Rightarrow G$	0.060000	0.666667
5	$H \Rightarrow I$	0.060000	0.428571
6	$A \Rightarrow C$	0.010000	0.100000
7	$BE \Rightarrow D$	0.010000	0.166667
8	$B \Rightarrow H$	0.010000	0.076923
9	$E \Rightarrow D$	0.030000	0.214286
10	$C \Rightarrow J$	0.020000	0.181818
11	$I \Rightarrow H$	0.060000	0.400000
12	$D \Rightarrow E$	0.030000	0.250000
13	$J \Rightarrow A$	0.050000	0.416667
14	$DJ \Rightarrow C$	0.010000	1.000000
15	$I \Rightarrow G$	0.030000	0.200000
16	$C \Rightarrow G$	0.020000	0.181818
17	$DF \Rightarrow E$	0.010000	1.000000
18	$J \Rightarrow C$	0.020000	0.166667
19	$G \Rightarrow F$	0.060000	0.375000
20	$B \Rightarrow G$	0.000000	0.000000
21	$I \Rightarrow A$	0.010000	0.066667
22	$D \Rightarrow C$	0.070000	0.583333
23	$E \Rightarrow B$	0.060000	0.428571

Table 4: Comparing the elite individual at each generation

Generation	The total number of extracted rules	Fitness
300	89	0.076441
500	92	0.078151
1000	89	0.080956

ing $(A \Rightarrow J)$, the records $(7 \sim 12)$ are for $(B \Rightarrow E)$, the records $(13 \sim 18)$ are for $(C \Rightarrow D)$, the records $(19 \sim 24)$ are for $(F \Rightarrow G)$, and the records $(25 \sim 30)$ are for $(H \Rightarrow I)$. The records from 31 to 100 are made randomly.

Table.2 shows *Support* and $Confidence$ of the given 5 kinds of rules and Table.3 shows the results with the proposed algorithm. Comparing Table.2 and Table.3, we can see that the rules from 1 to 5 in Table.2 and.3 are the same. The results from 6 to 23 in Table.3 denote the rules extracted from the records (31 to 100). These results show that it is possible to extract rules using the proposed algorithm.

Table.4 shows the total number of rules and the fitness from the elite individual at each generation. According to Table.4, we can see the elite individual extracted 92 rules at 500th generation, which are more than the ones of 1000th generation. However, we also see the fitness at 1000th generation is higher than 500th generation. We can see the elite individual at 1000th generation extracted better rules than the one at 500th generation according to the proposed algorithm because the fitness of 1000th generation is higher than that of 500th generation, although the elite individual at 500th generation extracted more rules than that at 1000th generation.

The simulation results show that *Support* and $Confidence$ of the extracted rules increase as generation goes on. This indicates that the proposed algorithm is useful to extract rules and can extract many effective rules through generations.

Fig.7 denotes the fitness of the elite individual, the average fitness values over all individuals, and the worst fitness at each generation. We extracted the rules of Table.3 from the elite individual.

6 Conclusions

We proposed the data mining technique using GNP in this paper. And we studied the effectiveness of the proposed method in terms of extracting rules. Concretely, the fitness of each individual is calculated using *Support* and $Confidence$ of the database, the better individuals having the higher fitness are selected, and finally the effective rules from the elite individual are extracted.

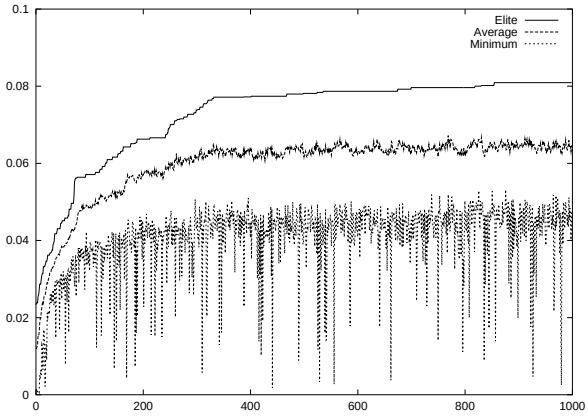

Figure 7: Simulation result

We have studied the effectiveness of the data mining using the proposed algorithm so far. The comparison of the proposed method with other methods is going underway.

References

[1] T. Fukuda, Y. Morimoto, and T. Tokuyama, "Data Mining - in japanese -", *Kyoritsu Press*, 2002.

[2] K. Hironobu, K. Hirasawa, J. Hu, J. Murata, M. Kosaka, "Network Structure Oriented Evolutionary Model : Genetic Network Programming", *Trans. of the Society of Instrument and Control Engineers, Vol.38, No.5, pp.485-494*, 2002.

[3] K. Hirasawa, M. Okubo, J. Hu, J. Murata, Y. Matsuya, "Co-evolution of Hetero Multiagent Systems using Genetic Network Programming", *IEEJ Trans. EIS, Vol.123, No.3, pp.554-551*, 2003.

[4] K. Hironobu, K. Hirasawa, J. Hu, J. Murata, "Variable Size Genetic Network Programming", *IEEJ Trans. EIS, Vol.123, No.1, pp.57-66*, 2003.

[5] David J. Hand, Heikki Mannila, Padhraic Smyth "Principles of Data Mining", *MIT Press*, 2001.