Extraction of Rules from Coronary Heart Disease Database Using Automatically Defined Groups

Akira Hara, Takumi Ichimura,
Tetsuyuki Takahama, and Yoshinori Isomichi

Faculty of Information Sciences, Hiroshima City University,
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
{ahara, ichimura, takahama, isomichi}@its.hiroshima-cu.ac.jp
http://www.chi.its.hiroshima-cu.ac.jp/

Abstract. Much of the research on extracting rules from a large amount of data has focused on the extraction of a general rule that satisfies as many data as possible. In the field of health care where people's lives are at stake, the exceptional rules for rare cases are also important. In this paper, we describe the knowledge acquisition from data containing such multiple rules. We consider that a multi-agent approach is effective for extracting multiple rules. In order to realize this approach, we propose a new method using an improved Genetic Programming method, Automatically Defined Groups (ADG). By using this method, the clustering of data is performed by sharing roles among agents, and each agent takes charge of rule extraction in the assigned data. As a result, all rules are extracted by multi-agent cooperation. We applied this method to coronary heart disease databases, and showed the effectiveness of this method.

1 Introduction

Recently, patient diagnostic data in hospitals have been accumulated in a database through the advance of information technology. Much research on extracting rules from a large amount of data has focused on the extraction of a general rule that satisfies as many data as possible. Some data which do not satisfy a general rule may be abandoned as exceptional data or noises. However, medical treatment is a field where people's lives are at stake. Therefore, knowledge that can be used for a small number of patients with exceptional symptoms is also very important. A key point to the achievement of high-quality medical treatment is whether such patients are treated appropriately without missing the symptoms.

In this paper, we extract rules from real medical data about coronary heart diseases. We aim to extract not only general diagnostic rules but also exceptional diagnostic rules. Moreover, we aim not only to improve the prediction accuracy but also to acquire the useful and comprehensible knowledge. In order to realize them, we use an improved Genetic Programming method for rule extraction.

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3214, pp. 1089–1096, 2004.

This method is Automatically Defined Groups (ADG)[2, 3]. Next, we describe this method.

2 Automatically Defined Groups

In this research, we handle the data containing multiple rules, and aim to do both the clustering of data and rule extraction from each cluster. We use a multi-agent approach to solve this problem. That is, the data are divided among agents. This corresponds to the clustering of data. And each agent generates a rule for the assigned data. This corresponds to the rule extraction in each cluster. As a result, all rules are extracted by multi-agent cooperation. In order to use this approach, however, we do not know the number of rules hidden in data and how to allot data to each agent. Moreover, as we prepare abundant agents, the number of tree structural programs in an individual increases. Therefore, the search performance becomes worse.

In order to solve the problems, we have proposed ADG. This is a method to optimize both the grouping of agents and the program of each group in the process of evolution. By grouping multiple agents, we can prevent the increase of search space and perform an efficient optimization. Moreover, we can easily analyze the agents' behavior. The acquired group structure is utilized for understanding how many roles are needed and which agents have the same role. That is, the following three points are automatically acquired by using ADG.

- How many groups (roles) are required to solve the problem?
- Which group does each agent belong to?
- What is the program of each group?

A team that consists of all agents is regarded as one GP individual. One GP individual maintains multiple trees, each of which functions as a specialized program for a distinct group. We define a group as the set of agents referring to the same tree for the determination of their actions. All agents belonging to the same group use the same program.

Generating an initial population, agents in each GP individual are divided into random groups. Basically, crossover operations are restricted to corresponding tree pairs. For example, a tree referred to by an agent 1 in a team breeds with a tree referred to by an agent 1 in another team. However, we consider the sets of agents that refer to the trees used for the crossover. The group structure is optimized by dividing or unifying the groups according to the relation of the sets. Individuals search solutions as their group structures gradually approach the optimal structure.

The concrete processes are as follows: We arbitrarily choose an agent to two parental individuals. A tree referred to by the agent in each individual is used for crossover. We use T and T' as expressions of these trees, respectively. In each parental individual, we decide a set $A(T)$, the set of agents that refer to the selected tree T. When we perform a crossover operation on trees T and T', there are the following three cases.

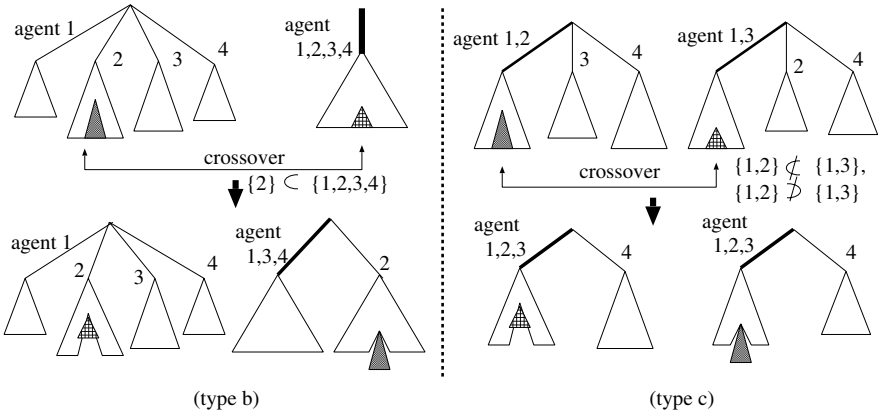

Fig. 1. Examples of crossover

(Type a) If the relation of the sets is $A(T) = A(T')$, the structure of each individual is unchanged.

(Type b) If the relation of the sets is $A(T) \supset A(T')$, the division of groups takes place in the individual with T, so that the only tree referred to by the agents in $A(T) \cap A(T')$ can be used for crossover. The individual which maintains T' is unchanged. Fig. 1 shows an example of this crossover.

(Type c) If the relation of the sets is $A(T) \not\supset A(T')$ and $A(T) \not\subset A(T')$, the unification of groups takes place in both individuals so that the agents in $A(T) \cup A(T')$ can refer to an identical tree. Fig. 1 shows an example of this crossover.

We expect that, by using this method, the search works efficiently and the adequate group structure is acquired. Besides, the acquired group structure becomes a clue for understanding the cooperative behavior and necessary division of labor.

3 Extracting Rules from Medical Database Using ADG

3.1 Coronary Heart Disease Database

In this section, we apply the proposed method to knowledge acquisition from real medical data. We use the database on coronary heart diseases[1]. Data in coronary heart diseases database are divided into two classes: non-coronary heart disease cases (non-CHD) and coronary heart disease cases (CHD). Each patient's disorder is diagnosed according to the results of eight test items. The eight items tested are Cholesterol (TC), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Left Ventricular Hypertrophy (LVH), Origin (ORIGIN), Education (EDUCATE), Smoking (TABACCO), and Drinking (ALCOHOL).

The original results of some test items are provided as the real values with various ranges. So, we normalize each value. We find the maximum value (max) and minimum value (min) on each item in training data set, and i-th item's value x_i is normalized to X_i as follows:

$$X_i = (x_i - min_i)/(max_i - min_i)$$

In this research, we intend to construct a diagnostic system which can classify data into the appropriate class based on these eight tests. We use four training data sets (Train_A, X, Y, Z) and a test database (Test).

3.2 How to Apply ADG to Coronary Heart Disease Database

In order to judge whether each data is regarded as CHD case, we will find logical expressions such that only the data in CHD cases should satisfy. The logical expression is made by the conjunction of multiple terms. Each term is the combination of a test item and the value which can be taken. The following expression is an example.

$$\text{Rule for CHD} : (TC > 0.51) \wedge (TC < 0.68) \wedge (DBP > 0.49)$$

In this case, the logical expression has to return false for non-CHD cases.

In medical field, the diagnoses are largely dependent on each doctor's experience. Therefore, the diagnostic rule is not necessarily represented by a single rule. Moreover, some data can be classified into different results, even if the results of the tests are the same. We apply ADG to the diagnoses of coronary heart diseases with consideration of this background.

We describe the detail of rule extraction for CHD cases. Multiple trees in an individual of ADG represent the respective logical expressions. Each data in the training set is input to all trees in the individual. Then, calculations are performed to determine whether the data satisfy each logical expression. As illustrated by data 2 in Fig.2, the input data is regarded as CHD case if even one among the multiple logical expressions in the individual returns true. In contrast, as illustrated by data 1 in Fig.2, the input data is not regarded as CHD case if all logical expressions in the individual return false.

The concept of each agent's load arises from the viewpoint of cooperative problem solving by multiple agents. The load is calculated from the adopted frequency of each group's rule and the number of agents in each group. The adopted frequency of each rule is counted when the rule successfully returns true for each CHD data. As illustrated by data 3 in Fig.2, if multiple trees return true for a CHD data, the tree with more agents is adopted. When the k-th agent belongs to group g, the load of the agent is defined as follows.

$$w_k = \frac{(\text{adopted frequency of } g) \times N_{agent}}{(\text{Number of agents which belong to } g) \times N_{all\_adoption}}$$

In this equation, N_{agent} represents the number of all agents in one GP individual, and $N_{all\_adoption}$ represents the sum of adopted frequencies of all groups.

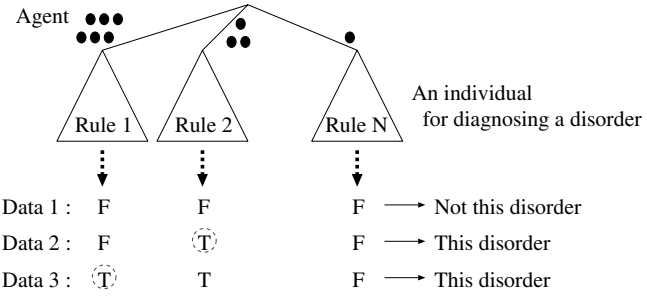

Fig. 2. Diagnostic system for a particular disorder

By balancing every agent's load, more agents are allotted to the group that has a greater frequency of adoption. On the other hand, the number of agents in the less adopted group becomes small. Therefore, we can acquire important knowledge about the ratio of use of each rule. The ratio indicates how general each rule is for judgment of the disorder. Moreover, when other cases are judged to be true through a mistake of a rule, it is thought that the number of agents who support the rule should be small.

To satisfy the requirements mentioned above, fitness f is calculated by the following equation. We maximize f by evolution.

$$f = -\frac{miss\_target\_data}{N_{CHD}} - \alpha \frac{misrecognition}{N_{nonCHD}} - \beta \frac{\sum_{N_{nonCHD}} fault\_agent}{misrecognition \times N_{agent}} - \delta V_w \tag{1}$$

In this equation, N_{CHD} and N_{nonCHD} represent the number of CHD cases and non-CHD cases in database respectively. $miss\_target\_data$ is the number of missing data in the target CHD data that should have been judged to be true. $misrecognition$ is the number of mistakes through which non-CHD data is regarded as CHD case. When the rule returns true for non-CHD data, $fault\_agent$ is the number of agents who support the wrong rule in each data. So, the third term represents the average rate of agents who support the wrong rules when misrecognition happens. V_w is the variance of every agent's load. In addition, in order to inhibit the redundant division of groups, f is multiplied by γ^{G-1} ($\gamma > 1$) according to the increase of the number of groups, G, in the individual.

By evolution, one of the multiple trees learns to return true for a data in the CHD cases, and all trees learn to return false for non-CHD cases. Moreover, agents are allotted to respective rules according to the adopted frequency, and the allotment to a rule with more misrecognition is restrained. Therefore, the rule with more agents is the typical and reliable diagnostic rule, and the rule with less agents is an exceptional rule for the rare case.

The following points are regarded as the advantages of ADG.

– ADG enables us to extract rules for exceptional data that is likely to be missed by a single rule.

- It is easy to judge by the number of agents whether the acquired rules are typical ones or exceptional ones.
- It is easy to understand the acquired rules, because typical rules and exceptional rules are clearly separated.

Table 1 shows GP functional and terminal symbols. We impose constraints on the combination of these symbols. Terminal symbols do not enter directly in the arguments of the `and` function. Test items such as `TC` enter only in arg0 of `gt` and `lt`. Real values enter only in arg1. Crossover and mutation that break the constraints are not performed.

The parameter settings of ADG are as follows: Population size is 500, crossover rate is 0.9, mutation rate per individual is 0.95, group mutation rate is 0.04, and the number of agents is 50.

4 Results

In this section, ADG is applied to the training data so that only CHD cases can satisfy the rules. We describe the detail of an experiment using Train_Z, which are consisted of 400 CHD cases, and 3600 non-CHD cases. The respective weights in equation(1) are $\alpha = 1.0$, $\beta = 0.0001$, $\delta = 0.01$, and $\gamma = 1.001$.

Fig. 3 shows the average group number by generation. The number of groups corresponds to the number of extracted rules. We can see from these figures that individuals are optimized as the number of necessary rules is searched.

As a result, 50 agents in the best individual are divided into 12 groups. We show the acquired rules that correspond to the tree structural programs in the best individual. Rules are arranged according to the number of agents that support each rule, and each terminal real value is transformed to original range. The rules with more agents are frequently adopted rules. The rules with less agents are rules for exceptional data.

Rule 1 (19 Agents): (SBP $>$ 179)
Rule 2 (7 Agents): (LVH $=$ 1)
Rule 3 (6 Agents): (TC $>$ 199) $\wedge$ (SBP $>$ 141) $\wedge$ (DBP $>$ 99) $\wedge$ (DBP $<$ 112) $\wedge$ (LVH $=$ 0) $\wedge$ (EDUCATE $<$ 3) $\wedge$ (ALCOHOL $<$ 34.54)
Rule 4 (6 Agents): (TC $>$ 264) $\wedge$ (SBP $>$ 150) $\wedge$ (TABACCO $>$ 1) $\wedge$ (ALCOHOL $<$ 44.9)
Rule 5 (2 Agents): (TC $>$ 168) $\wedge$ (TC $<$ 252) $\wedge$ (SBP $>$ 127) $\wedge$ (DBP $>$ 106) $\wedge$ (TABACCO $>$ 2) $\wedge$ (ALCOHOL $>$ 19.0)
Rule 6 (2 Agents): (TC $>$ 310)
Rule 7 (2 Agents): (SBP $>$ 141) $\wedge$ (DBP $>$ 104) $\wedge$ (LVH $=$ 0) $\wedge$ (EDUCATE $<$ 2) $\wedge$ (TABACCO $>$ 0) $\wedge$ (TABACCO $<$ 3)
Rule 8 (2 Agents): (TC $>$ 242) $\wedge$ (TC $<$ 296) $\wedge$ (DBP $>$ 109) $\wedge$ (ORIGIN $=$ 1) $\wedge$ (TABACCO $>$ 0) $\wedge$ (ALCOHOL $>$ 15.9)
Rule 9 (1 Agents): (TC $>$ 214) $\wedge$ (SBP $>$ 152) $\wedge$ (DBP $>$ 85) $\wedge$ (EDUCATE $<$ 1) $\wedge$ (TABACCO $<$ 2)
Rule 10 (1 Agents): (DBP $>$ 79) $\wedge$ (DBP $<$ 84) $\wedge$ (ALCOHOL $>$ 37.5)
Rule 11 (1 Agents): (TC $>$ 233) $\wedge$ (SBP $>$ 160) $\wedge$ (DBP $>$ 98) $\wedge$ (DBP $<$ 132) $\wedge$ (ORIGIN $=$ 0) $\wedge$ (EDUCATE $<$ 3) $\wedge$ (ALCOHOL $<$ 35.1)

Table 1. GP Functions and Terminals

Symbol	#args	functions
`and`	2	arg0 ∧ arg1
`gt`	2	if (arg0 > arg1) return T else return F
`lt`	2	if (arg0 < arg1) return T else return F
`TC`, `SBP`, ...	0	normalized test value
`0.0` – `1.0`	0	real value

Table 2. Recognition rates

Dataset	recognition rate
Train_A	70.0% (67.8%)
Train_X	70.2% (68.5%)
Train_Y	70.1% (68.6%)
Train_Z	75.0% (66.6%)

Rule 12 (1 Agents): (TC > 186) ∧ (TC < 330) ∧ (SBP > 169) ∧ (DBP > 99) ∧ (DBP < 114) ∧ (LVH = 0) ∧ (TABACCO > 0) ∧ (TABACCO < 3) ∧ (ALCOHOL < 34.5)

The judgment accuracy for 4000 training data is as follows. One or more rules return true for 308 cases of 400 CHD cases, and all rules successfully return false for 2691 of 3600 non-CHD cases. Therefore, the recognition rate to the training data is 75.0%.

We examined which rule's output is adopted for the 308 successful data. The counts of adoption of these twelve rules are 115, 46, 38, 36, 16, 13, 12, 10, 9, 7, 4, and 2 times, respectively. These data result from the effects of the third and fourth terms of the fitness equation (1). The ratio of adopted frequencies of the respective rules does not completely correspond to the ratio of agents in each group, because there is a requirement to reduce the number of agents who support the rule with misrecognition data. However, the rule with more agents tends to have a higher adopted frequency. Both typical rules for frequent cases and exceptional rules for rare cases were extracted successfully. Moreover, this system was applied to 13000 test data. As a result, it succeeded in the classification of 8655 cases. The recognition rate was 66.6%.

We also applied this method to other training data sets (Train_A, X, Y), and examined the performance of each result for both training and test data. Table 2 shows the recognition rates. The parenthetic values in the table mean the recognition rates for Test dataset. The acquired rules are represented by simple logical expressions. So, we can easily acquire diagnostic knowledge from the rules. However, constraint of the expressions may have a bad influence upon the recognition rate. By modifying the GP symbols so that the rules can represent more complex expressions (*e.g.* DBP > 1.2SBP etc.), we have to improve the recognition rate with keeping the comprehensibility.

5 Conclusions and Future Work

In this research, we proposed a new method using ADG for the purpose of the extraction of multiple rules. In this method, the clustering of data and rule

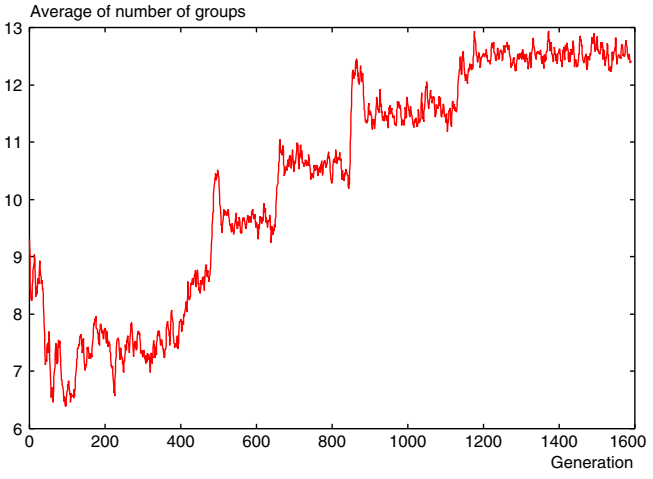

Fig. 3. Change of the average of the number of groups

extraction in each cluster are performed simultaneously. We showed the effectiveness of this method by the application to medical data.

The diagnostic rules were extracted mechanically from only numerical data. Some rules may be not accepted easily in the field of medicine, because they may include absurd combinations of items besides the common sense of doctors. By taking knowledge of medical treatment into account during the process of optimization, more effective rules can be acquired. The optimization by ADG using such knowledge is a future work. In addition, we have to investigate the usefulness of extracted rules from the viewpoint of health care.

Acknowledgments

This work was supported by a Grant-in-Aid for Young Scientists (B) (No. 15700199) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. In addition, this was also supported by a Hiroshima City University Grant for Special Academic Research (General Study).

References

1. M. Suka, T. Ichimura and K. Yoshida: "Development of Coronary Heart Disease Database", *Proc. The Eighth Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'2004)* (to appear)
2. A. Hara and T. Nagao: "Emergence of cooperative behavior using ADG; Automatically Defined Groups", *Proc. The Genetic and Evolutionary Computation Conference 1999*, pp.1039-1046 (1999)
3. A. Hara, T. Ichimura, T. Takahama and Y. Isomichi: "Extraction of rules by Heterogeneous Agents Using Automatically Defined Groups", *Proc. The Seventh Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'2003)*, Vol.2, pp.1405-1411 (2003)