Constructing an Initial Knowledge Base for Medical Domain Expert System using Induct RDR

Jonghwan Hyeon [1], Kyo-Joong Oh [1], You Jin Kim [1], Hyunsuk Chung [2], Byeong Ho Kang [2], and Ho-Jin Choi [1]
School of Computing, KAIST, Daejeon, Republic of Korea [1]
School of Engineering and ICT, University of Tasmania, Hobart, Tasmania, Australia [2]
{hyeon0145, aomaru, 117kyjin, hojinc}@kaist.ac.kr, {David.Chung, Byeong.Kang }@utas.edu.au

***Abstract*—This paper describes how we build an initial knowledge-base of ripple-down rules (RDR) in medical domain. In medical domain, all decisions are made by the domain experts. Increasing a complexity of disease and various symptoms, there are some attempts to introduce an expert system in medical domain these days. To construct the expert system, it needs to extract the expert's knowledge. To do that, we use ripple-down rules (RDR) which allows experts to modify their knowledge base directly because it provides a systematic approach to do that. We also use Induct RDR which builds a knowledge base from existing data to reduce experts' burden of adding their knowledge from the bottom up. The expert system should produce multiple comments from a test set, which is multiple classification problem. However, Induct RDR only deals with a single classification problem. To handle this problem, we divide a test set into 18 categories which is almost the single classification problem and apply Induct RDR to each category independently. Using this approach, we can improve the missing rate about 70% compared to an approach not dividing into several categories.**

Keywords—knowledge base; medical domain; ripple-down rules; induct RDR; multiple classification problem

I. INTRODUCTION

In medical domain, all decisions are made by the domain experts. All decisions in medical domain are directly related to patients' life, and wrong decisions can give a huge damage to patients. Also recently increasing a complexity of disease and various symptoms, it is being hard for a small number of medical experts to observe a variety of cases of patients and diagnose a disease that they have. So there are some attempts to introduce an expert system in medical domain these days due to time reasons and a lack of medical experts.

To construct the expert system, it needs to extract the expert's knowledge as a formation that the system can understand. In this step, the important thing is how easy the maintenance will be. Previous works have needed continuous interventions of knowledge engineers to reflect new knowledge after constructing the knowledge base. However, ripple-down rules (RDR) allows experts to modify their knowledge base directly because it provides a systematic approach to do that without any interventions of knowledge engineers. [1]

When constructing the expert system, it is common best practice to build an initial knowledge base from existing data using a statistical method rather than adding experts' knowledge from the bottom up. If not, it could be a big burden to experts because they need to add all their knowledge to the base. To handle this problem, we use Induct RDR which builds a knowledge base from existing data. [2]

In this paper, we construct the initial knowledge base that will be used for the expert system from existing data using Induct RDR. The contribution of this work is that we utilize the Induct RDR which is for the single classification problem into the multiple classification problem.

The rest of this paper is organized as follows. Section 2 provides some background for understanding and related works. Section 3 gives our methods and approaches in detail. Section 4 describes the experiments and results on constructing the initial knowledge base. Finally, section 5 provides conclusions of this paper and future works.

II. BACKGROUND

RDR is a systematic approach to acquire knowledge from experts. In RDR, we infer conclusions using RDR base which is n-ary tree. From the root node, if the condition of current node is satisfied, we go further to next level. Otherwise, we stop and produce conclusions on finally fired nodes. In RDR, all modification of RDR base is done by experts using the systematic approach. After inferencing the input, the experts should verify the output and they can modify rules if conclusions are wrong. [1]

Parallel distributed processing (PDP) diagnostic system is a medical expert system that diagnoses a disease based on patients' symptom. This system consists of 3 layers. In the input layer, each node indicates an item of a survey asking patients' symptom. Each node in output layer indicates each disease and has a value between 0 and 1 which means a probability that the patient has this disease. In PDP diagnostic system, there is a threshold. So if a node in output layer exceeds the threshold, we can conclude that this patient has that disease. Because the output layer has many nodes, the PDP diagnostic system can infer multiple conclusions. [3]

PRIMERSOSE-REX is a medical expert system based on the rough set theory which is a kind of the inductive learning. This system can extract not only a classification rule but also additional information that needs in diagnosis. This system produces a rule-based headache and facial pain Information system (RHINOS) as a result. [4]

eFilter system is a system to recommend a diet menu considering a patient's health. This system is based on case-based approach and using RDR as a decision system. Domain experts find a similarity among cases using their experts' knowledge and intuition. However, it is hard to provide knowledge engineers with rules related to the above process. Case-based reasoning provides a solution to this kind of problem. [5]

Almost all expert systems have problems on analysis and maintenance. To deal with this problem, Han et al. use two approaches. To overcome analysis problem, they use an agile software development approach. Also to handle maintenance problem, they use the RDR as the expert system development. [6]

III. Methods and Approaches

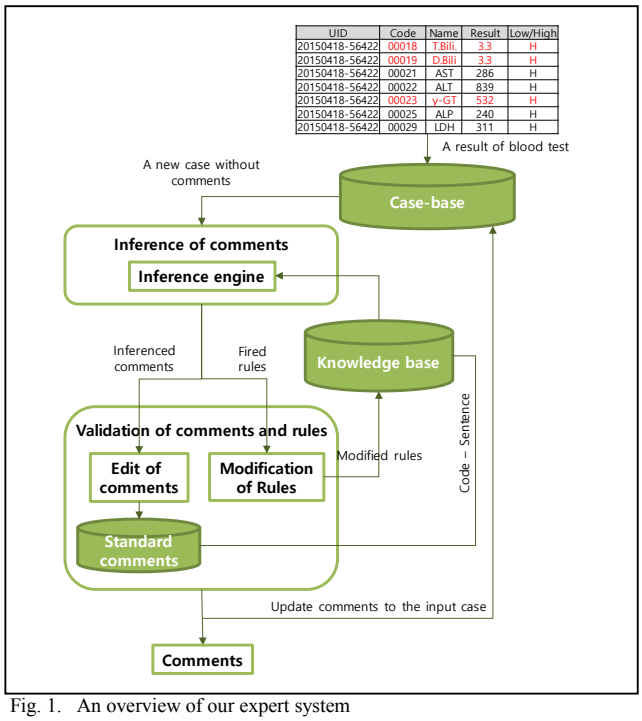

Fig. 1. An overview of our expert system

A. Ripple-down Rules (RDR)

We plan to build the expert system that produces comments from test data. Because all test data are numerical in the medical domain and all decisions are made by domain experts according to reference data, we think that this process is very suitable for a concept of the expert system. Therefore, we decided to build the expert system that can substitute experts' decision making process.

In the expert system, it is very important to decide which decision systems we will use because the decision system affect all processes from how to build an initial knowledge base to how to maintain the knowledge base. We decide to use RDR as the decision system in our expert system because it can minimize interventions of knowledge engineers after constructing the initial knowledge base. Previous works have needed continuous interventions of knowledge engineers to reflect new knowledge and that is the reason why previous expert systems have not worked well. However, RDR provides the systematic approach to modify their knowledge base. So, they can reflect their new knowledge easily to the base. Thus, they don't need any interventions of knowledge engineers and can construct the base incrementally.

We also build the initial knowledge base using Induct RDR which builds a knowledge base from existing data rather than build the base from the bottom up with experts. We already have a fluent of data that is populated by experts. If we can construct the initial base utilizing existing data, we can reduce a burden of experts to adding their knowledge because they only need to add missing parts and modify wrong parts in the base.

B. Multiple Classification

The expert system to infer comments should produce multiple comments from a test set. In other words, this is a multiple classification problem which takes an input and produces multiple outputs. However, previous works including Induct RDR are dealing with a single classification problem which takes an input and produces an output. To handle this problem, therefore, we divide a test set into 18 categories.

TABLE I. 18 Categories

Category	Category
Anemia	Liver
Blood	Pancreas
Blood sugar	Rheumarthritis
Blood type	Stool
Electrolyte	Syphilis
Hepatitis Virus	Thyroid
Infection	Tumor
Kidney	Urine
Lipid	Etc.

In these categories, almost all categories satisfy the single classification problem. However, some domains such as hepatitis virus and etc. cannot satisfy the single classification problem. For example, hepatitis virus needs several conclusions in different situations like hepatitis A virus, hepatitis B virus and so on. To deal with this problem, we duplicate one test set so that they all have one conclusion when generating data.

TABLE II. An Approach To Handle Multi-Classification

G	A	…	C
93	13		237
			132
			112

⇒

G	A	…	C
93	13		237
93	13		132
93	13		112

[a] G: Glucose, A: AST, C: Conclusion

After generating data in this way, we apply Induct RDR to each category independently.

IV. Experiments and Results

We received anonymized test data set from Seegene (seegene.co.kr) which is a medical foundation. Each test set contains several tests. Each test consists of its code, its result in numeric, its low limit and its high limit. And all test sets have comments according to the test results from experts. We aim to build the expert system which can infer comments from a test set that does not have any comments.

We firstly preprocess test sets. There are many test codes which have same test contents. For example, the test code 00011 and 00530 indicate a value of Glucose. So, we map various test codes to one test item. And rather than dealing with numerical data, we quantize a test result as low, normal and high using provided low and high limits.

Also as we said, we divide test sets into 18 categories to handle this multiple classification problem as the single classification problem. We apply Induct RDR on the test set divided into 18 categories. Because each category almost has the single classification domain, it is so suitable for Induct RDR. After building the initial knowledge base, we infer comments from the test set using each knowledge base independently. And we compare these inferred comments with the existing comments by experts

```
Root THEN 187
  [1] IF (GTP == hc) THEN 182
    [2] IF (AST == hc) & (ALT == hc) THEN 69
      [3] IF (LDH == hc) & (ALP == nc) THEN 65
        [4] IF (Cholinesterase == nc) THEN 63
      [5] IF (BilirubinTotal == hc) & (LDH == nc) THEN 68
      [6] IF (LDH == hc) & (BilirubinTotal == hc) THEN 60
      [7] IF (LDH == hc) & (BilirubinIndirect == NA) THEN 61
      [8] IF (LAP == NA) & (ALP == hc) THEN 53
      [9] IF (Cholinesterase == NA) & (BilirubinDirect == hc) THEN 68
      [10] IF (LAP == hc) & (ALP == nc) THEN 57
      …
```

Fig. 2. Example of a knowledge base of the liver cateogry

We use 7610 test set to populate the initial knowledge using Induct RDR and evaluate the base using 1000 test set. The test results are summarized on the Table III. We investigate a missing rate which means how many comments are missed in inferenced comments by the expert system. As a result, we can get 46.91% missing rate. This is 70% improved result compared to a previous approach before we divide test set into several categories. However, we get 96.01% an additional rate which means how many comments are added in inferenced comments by the expert system. We think this is because we always generate at least 18 comments due to our approach.

TABLE III. Comparison With Our Approach

	Previous	Our Approach
Missing Rate	66.22%	46.91%
Additional Rate	21.61%	96.01%

V. Conclusion

We construct the expert system in medical domain. We firstly preprocess test dataset and use Induct RDR to build the initial knowledge base. However, its performance is not good because Induct RDR is based on the single classification problem. So we divide the test dataset into 18 categories which are almost the single classification problem. After that, we can improve the missing rate about 70%. But the performance is still not good. So we are planning for continuing research to automatically build multiple classification RDR from data.

Acknowledgment

This work was supported by the Industrial Strategic Technology Development Program, 10052955, Experiential Knowledge Platform Development Research for the Acquisition and Utilization of Field Expert Knowledge, funded by the Ministry of Trade, Industry & Energy (MI, Korea)

References

[1] Compton, P., Edwards, G., Kang, B., Lazarus, L., Malor, R., Menzies, T., ... & Sammut, C. (1991). Ripple down rules: possibilities and limitations. In Proceedings of the Sixth AAAI Knowledge Acquisition for Knowledge-Based Systems Workshop, Calgary, Canada, University of Calgary (pp. 6-1).

[2] Gaines, Brian R., and Paul Compton. "Induction of ripple-down rules applied to modeling large databases." Journal of Intelligent Information Systems 5.3 (1995): 211-228.

[3] Saito, K., & Nakano, R. (1988, July). Medical diagnostic expert system based on PDP model. In Neural Networks, 1988., IEEE International Conference on (pp. 255-262). IEEE.

[4] Tsumoto, S. (1998). Automated extraction of medical expert system rules from clinical databases based on rough set theory. Information sciences, 112(1), 67-84.

[5] Kovasznai, G. (2011, May). Developing an expert system for diet recommendation. In Applied Computational Intelligence and Informatics (SACI), 2011 6th IEEE International Symposium on (pp. 505-509). IEEE.

[6] Han, S. C., Yoon, H. G., Kang, B. H., & Park, S. B. (2014). Using MCRDR based Agile approach for expert system development. Computing, 96(9), 897-908.