From educational data mining model to the automated knowledge based system construction

2015 8th International Conference on Ubi-Media Computing (UMEDIA): 186 - 191

From Educational Data Mining Model to the Automated Knowledge Based System Construction

Kittisak Kerdprasop and Nittaya Kerdprasop
Knowledge Engineering and Data Engineering Research Units,
School of Computer Engineering, Suranaree University of Technology,
111 University Avenue, Nakhon Ratchasima 30000, Thailand.
{kerdpras, nittaya}@sut.ac.th

***Abstract*—A knowledge based system (KBS) has its advantage over conventional database systems in that it has the inference ability to deduce implicit knowledge from the explicitly stored information. KBS is however known to be labor intensive in its construction in the knowledge acquisition and elicitation phase. Researchers have tried to overcome this hindrance for more than four decades. Automatic creation of a knowledge base (KB) content is still a research topic of interest. In this paper, we propose the design of a framework that not only automatically creates a KB, but also constructs the inference and reasoning engine of the KBS. The KB content is elicited and transferred from the data mining model, whereas the engine (or shell) of the KBS is created from the decision rules. We demonstrate a case study in student loan payment decision using the visualized tools KNIME and WIN-PROLOG to generate a data mining model and a KBS shell, respectively.**

Keywords—automated knowledge base construction; knowledge based system; data mining model; decision tree; visualized tools

I. INTRODUCTION

Knowledge based system (KBS, also called expert system) is a broad term referring to the special kind of a computer program that works with information stored in the knowledge base. Knowledge base (KB) contains the explicit knowledge of human experts in a specific domain encoded in a machine-readable format. An intelligent computer program to infer implicit information from the stored knowledge with reasoning mechanism is called the inference engine. The separation of KB from the inference engine is the main characteristic of the KBS that makes it different from other conventional computer programs that integrate data and computer code in a single module.

Such KB-inference separation allows KB contents to evolve over times due to the addition of new knowledge, removal of obsolete information, or the change to some pieces of information. These actions regarding the KB contents can be done without any alteration to the inference engine. In fact the same inference and reasoning program can be applied to work with the KB of several domains. This well-organized architecture can save the development time of the new KBS because only the KB part has to be created, while the inference engine can be the reusable program.

The novel important concept of KB-inference separation has been introduced since the 1960s in the DENDRAL project by the Stanford heuristic programming team [14]. The DENDRAL system is the first practical knowledge-driven program applied to the medical and life science field. Since the 1980s the KBSs have shifted from the medical domain to various areas such as manufacturing [3], civil engineering [1], product design [5], and education [19].

Despite the extensive use of KBS in numerous application areas, building a new system is time-consuming due to the fact that knowledge acquisition and elicitation are still the labor-intensive tasks. These tasks require a knowledge engineer to collect and translate human experts' knowledge into the machine-readable representations [20]. Modern KBS development process has thus moved toward the automating methodology by applying intelligent knowledge extraction techniques. Such intelligent methods can be achieved through the machine learning and data mining technologies. There have been increasing number of research work attempting to apply learning techniques to automatically extract and elicit knowledge from databases [2], [12], [17], [22].

Our research takes the same direction as most current research work in an attempt to automate knowledge extraction and elicitation with learning facility of the data mining technology. The research work presented in this paper, however, moves the learning step towards a full scale of KBS development. Not only mining knowledge from existing data and information, we also design the knowledge transfer as a set of rules to augment the KB contents and develop in an automatic manner the inference and reasoning mechanism through the logic programming approach. Demonstration of the proposed method is provided in this paper via the generation of a KBS from the student loan payment data. Even though the case study is in the educational setting, the proposed idea of automated KBS construction is such a generalized framework that it can be applied to any application domains.

II. LITERATURE REVIEW

Automatic KB generation has long been a major concern of most knowledge engineers and computer scientists for more than twenty years. Learning knowledge via data mining approach and other artificial intelligence techniques has been studied extensively. Yoon and colleagues [23] proposed to use

978-1-4673-8270-0/15/$31.00 ©2015 IEEE

back-propagation learning algorithm to automatically generate a KB from patients' records. Holmes and Cunningham [9] discussed the design of the Explora data mining system to construct KB from the database records. Their system is intended to diagnose skin disease from the observed symptoms. Huang and Jensen [10] demonstrated the use of decision tree induction algorithm to automatically generate production rules for the remote sensing image analysis system. Zhu and teammates [24] illustrated the automated knowledge extraction using the Apriori association rule mining with the main purpose to support clinical decision model construction. Chorbev and colleagues [6] also studied the medical expert system that applies the rule induction algorithm to create knowledge through the Web based environment. Riccucci et al. [19] applied the inductive logic programming technique to acquire association rules from the deductive databases to support the intelligent tutoring system. Wang et al. [21] demonstrated the automatic construction of KB under some uncertainty conditions. Recent work of Clark et al. [7] proposed to construct KB automatically from text, instead of the database records. Mayfield and his research group [16] developed a tool called KELVIN to automatically create KB from text.

Most of the work appeared in the literature discussed and demonstrated mainly the use of data mining and machine learning techniques to automatically learn knowledge from the stored data. All the proposed methods can more or less alleviate the bottleneck problem of knowledge engineering process. But these research work has emphasized on the knowledge learning phase. The subsequent part of knowledge transferring to the KBS is normally intact. The work of Gaines and Shaw [8] is the most complete one in the sense that they discussed the idea of knowledge learning with the decision tree induction algorithm, the knowledge transformation to the appropriate from to be used later by inference engine, and the knowledge consultation through the expert system shell.

Our research objective resembles the work of Gaines and Shaw [8], but we learn the knowledge from a set of relations instead of a single file of dataset representing one relation. We also propose to use a visualized technique to learn knowledge and to generate inference engine for the KBS. Our proposed framework is presented in the next section.

III. A Framework of KBS Construction

The design of automatic KBS creation from the data mining model is presented in Fig. 1. The automated KBS construction framework is composed of three main modules: knowledge elicitation, knowledge transferring, and knowledge consultation. The first step of knowledge elicitation is relation integration and imputation. This is essentially a data integration phase of the data mining process. However, in our framework the data integration idea has been generalized to handle heterogeneous schemas of data. Some incomplete data items have to be imputed to make a single complete set of related data. This dataset is then used as a training data to generate a data mining model. If the learning objective is to learn some classification model based on a specific target, that target attribute has to be identified prior to the model induction step. The model is thus the output of knowledge elicitation module.

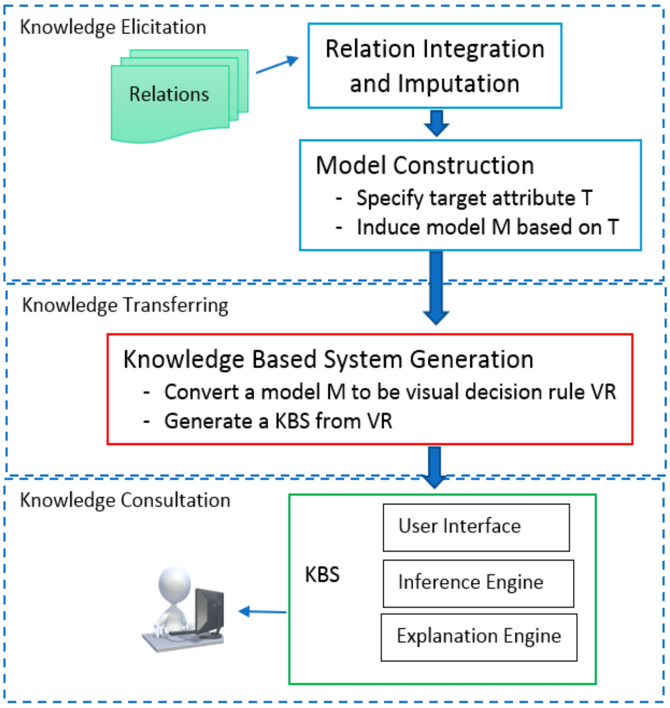

Knowledge Elicitation
Relations
Relation Integration and Imputation
Model Construction
- Specify target attribute T
- Induce model M based on T
Knowledge Transferring
Knowledge Based System Generation
- Convert a model M to be visual decision rule VR
- Generate a KBS from VR
Knowledge Consultation
KBS
User Interface
Inference Engine
Explanation Engine

Fig. 1. A framework of KBS construction from relations.

The knowledge transferring module is responsible for transforming the knowledge representation to be in a suitable format for the next stage of KBS construction. In our framework, the learned knowledge (i.e., the data mining model in a form of decision tree) has to be transformed as a decision rule set via a visualized tool (the example will be provided in the next section). Once the visual decision rule set has been correctly created, the KBS can be easily constructed through a logic programming tool. The automatically generated KBS has the user interface to interact with users via a GUI, the inference engine to deduce implicit knowledge and convey new information to the user, and the explanation engine to describe answer that has been given to user. The running example of this proposed framework is illustrated in the next section.

IV. Student Loan Payment Decision: A Case Study of Automated KBS Construction

To demonstrate the working example of our automated KBS, we use the student loan payment decision as a case study. The dataset had been used by Pazzani and Brunk [18] in 1991 to explain the idea of concept learning in their rule-based expert system named KR-FOCL. This dataset is available in the UCI machine learning repository [13]. We transform the relations in this dataset and impute the missing ones and then consolidate all relations into a single file. This data file is used as a training data to generate model through the KNIME information miner tool [4], [11]. The model is then transformed to be a decision rule in a graphical format of the VisiRule in the WIN-PROLOG software [15]. The KBS can finally be generated from the VisiRule. Details of the dataset and operations in each step can be explained as in the following subsections.

A. *Student Loan Payment Dataset*

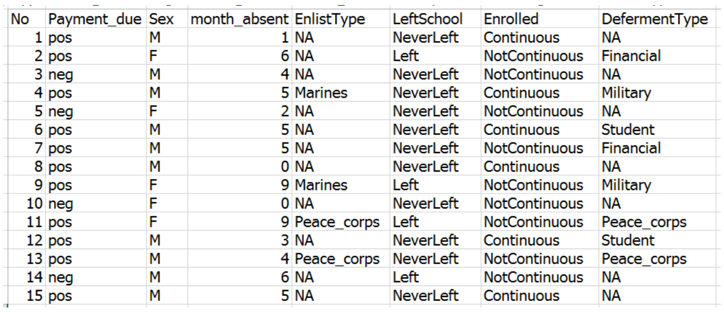
The objective of this dataset is to learn from individual records the general concept (or model) that can be used to predict whether a student is required to pay back an educational loan. Some students under specific conditions are exempt from the payment. There are 1000 examples (or cases) in this dataset: 643 cases are labelled as positive examples, 357 cases are negative examples. Positive cases are those students who have no payment due, whereas negative cases are those who have a loan payment due. The original goal of Pazzani and Brunk [18] who donated this dataset was to learn only the positive concept regarding those who are not required to pay a student loan (relation no_payment_due). The learned concept is in the format of a logic program. The positive examples, negative examples, and other ten auxiliary relations are also in a logic program formats (some of them are shown in Fig. 2).

```
no_payment_due(student10,neg).
no_payment_due(student5,neg).
no_payment_due(student3,neg).
…

disabled(student92).
disabled(student83).
disabled(student68).
…

filed_for_bankrupcy(student999).
filed_for_bankrupcy(student998).
filed_for_bankrupcy(student958).
…

male(student56).
male(student53).
male(student50).
…

enrolled(student10,ucb,2).
enrolled(student18,ucsd,4).
enrolled(student990,smc,10).
…

unemployed(student28).
unemployed(student22).
unemployed(student9).
…

enlist(student905,air_force).
enlist(student902,army).
enlist(student895,fire_department).
enlist(student894,marines).
enlist(student889,foreign_legion).
…
```

Fig. 2. Some relations in a student loan payment dataset.

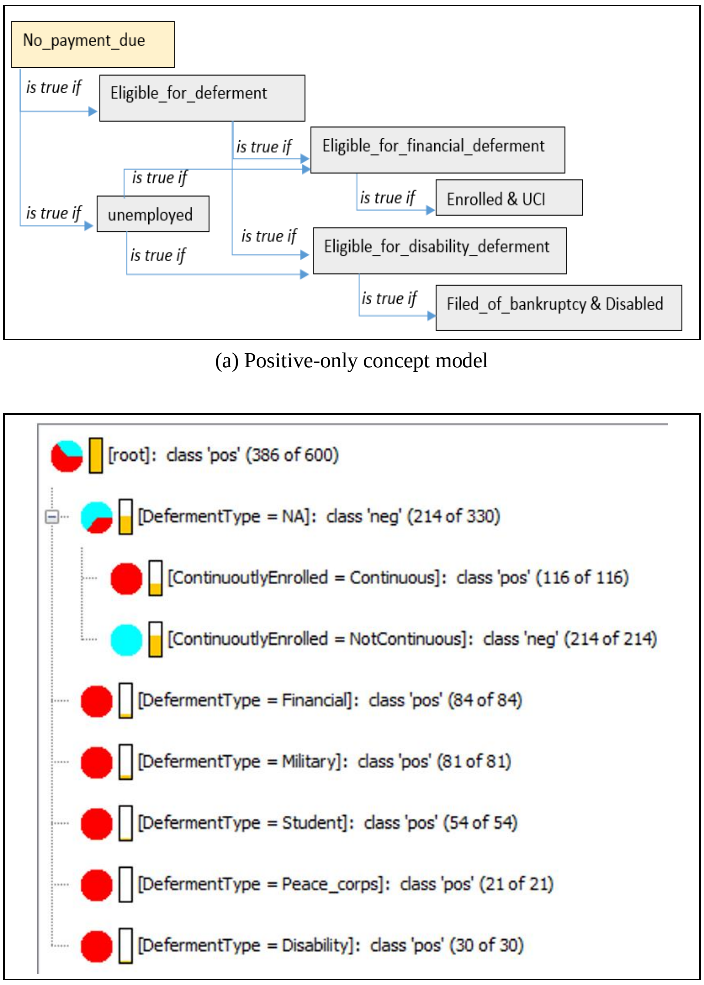
B. *Knowledge Elicitation Phase*

At the early stage of knowledge elicitation, we combine the 12 relations of a student loan dataset to be a single table. Some missing relation, such as a person who is female, is created. The transformed data are shown in Fig. 3. We then use the visual KNIME tool to induce a decision tree with the positive/negative class as a target attribute for tree induction. In the model induction phase, we use 600 cases as training data and the rest 400 cases are for testing. With the test data, the model is assessed to be 100% correct. The induced model is presented in Fig. 4. In Fig. 4(a), the original concept learned from the KR-FOCL of Pazzani and Brunk [18] is also displayed for a comparison to the one induced by ours (Fig. 4(b)).



No	Payment_due	Sex	month_absent	EnlistType	LeftSchool	Enrolled	DefermentType
1	pos	M	1	NA	NeverLeft	Continuous	NA
2	pos	F	6	NA	Left	NotContinuous	Financial
3	neg	M	4	NA	NeverLeft	NotContinuous	NA
4	pos	M	5	Marines	NeverLeft	Continuous	Military
5	neg	F	2	NA	NeverLeft	NotContinuous	NA
6	pos	M	5	NA	NeverLeft	Continuous	Student
7	pos	M	5	NA	NeverLeft	NotContinuous	Financial
8	pos	M	0	NA	NeverLeft	Continuous	NA
9	pos	F	9	Marines	Left	NotContinuous	Military
10	neg	F	0	NA	NeverLeft	NotContinuous	NA
11	pos	F	9	Peace_corps	Left	NotContinuous	Peace_corps
12	pos	M	3	NA	NeverLeft	Continuous	Student
13	pos	M	4	Peace_corps	NeverLeft	NotContinuous	Peace_corps
14	neg	M	6	NA	Left	NotContinuous	NA
15	pos	M	5	NA	NeverLeft	Continuous	NA

Fig. 3. Student loan payment data as a single dataset.



(a) Positive-only concept model

```
[root]: class 'pos' (386 of 600)
  [DefermentType = NA]: class 'neg' (214 of 330)
    [ContinuoutlyEnrolled = Continuous]: class 'pos' (116 of 116)
    [ContinuoutlyEnrolled = NotContinuous]: class 'neg' (214 of 214)
  [DefermentType = Financial]: class 'pos' (84 of 84)
  [DefermentType = Military]: class 'pos' (81 of 81)
  [DefermentType = Student]: class 'pos' (54 of 54)
  [DefermentType = Peace_corps]: class 'pos' (21 of 21)
  [DefermentType = Disability]: class 'pos' (30 of 30)
```

(b) Positive/negative concept model

Fig. 4. Student loan payment model (a) original model of Pazzani and Brunk [18], (b) a model induced by our method.

C. *Knowledge Transferring Phase*

A knowledge in a form of decision tree model as exhibited in Fig. 4(b) is then transferred to the next stage by change its representation as a decision rule. In this research, we apply a VisiRule tool of the WIN-PROLOG software [15]. A visualized decision rule is displayed in Fig. 5. If the diagram in the VisiRule tool is correctly created, the Prolog code for a KBS can then automatically generated (as in Fig. 6).

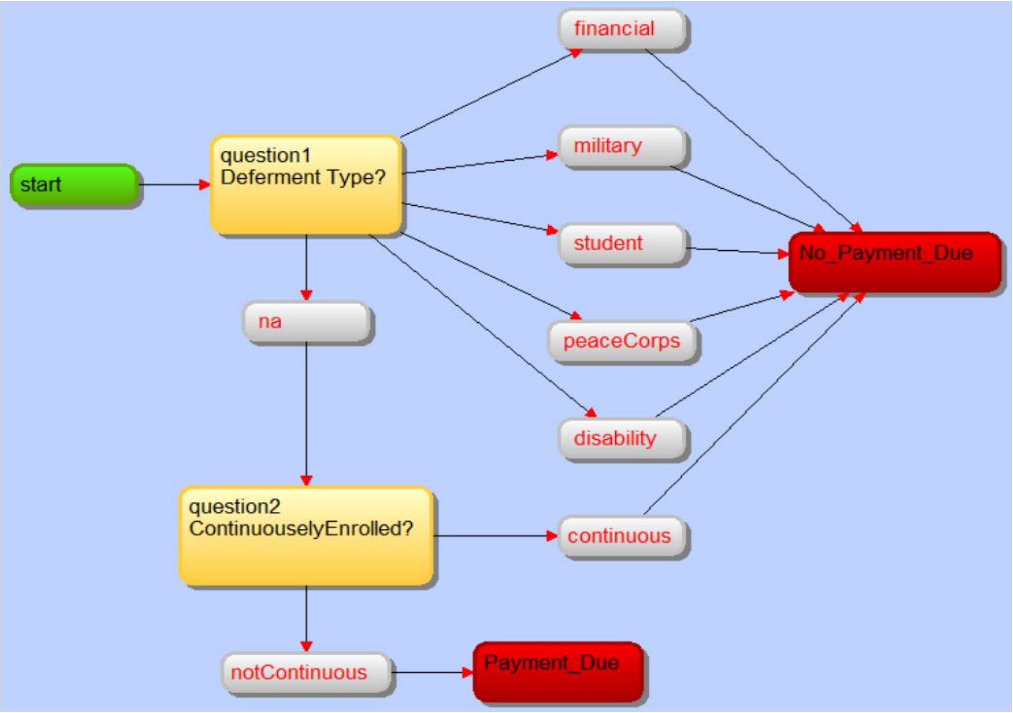
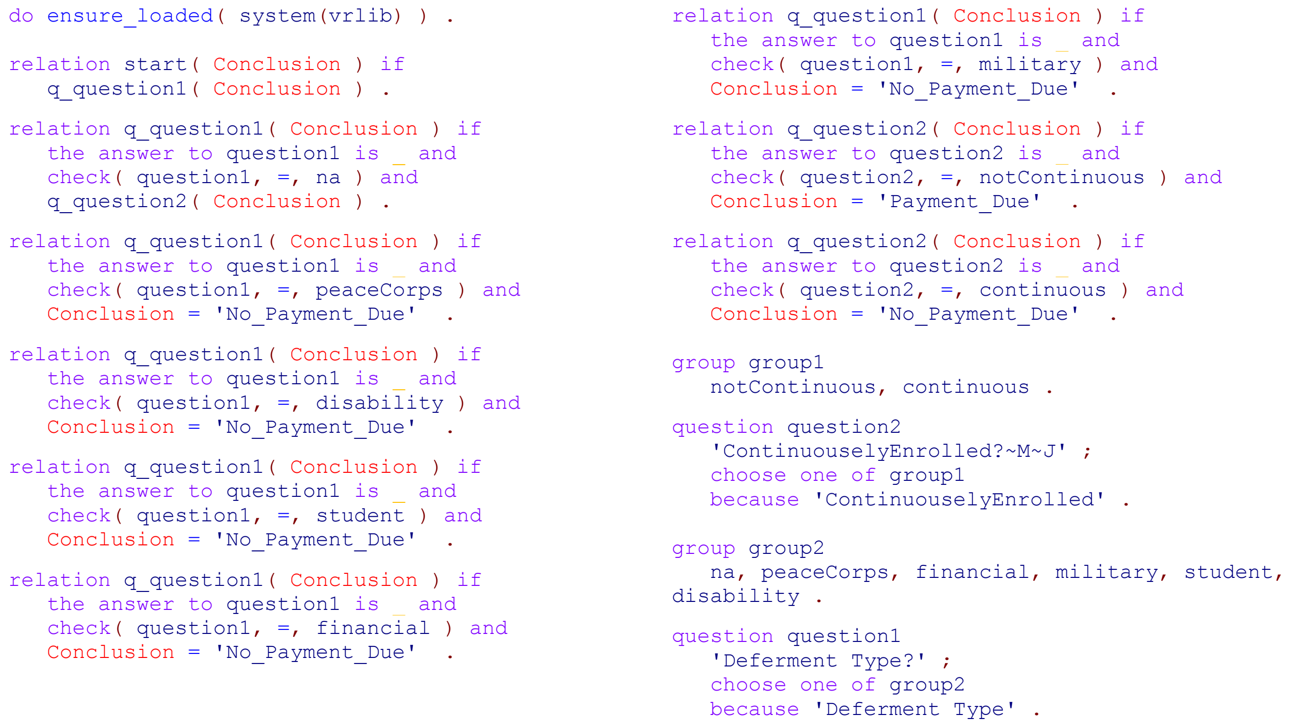


Fig. 5. A decision rule set in VisiRule to model a student loan payment decision.



```
do ensure_loaded( system(vrlib) ) .

relation start( Conclusion ) if
   q_question1( Conclusion ) .

relation q_question1( Conclusion ) if
   the answer to question1 is _ and
   check( question1, =, na ) and
   q_question2( Conclusion ) .

relation q_question1( Conclusion ) if
   the answer to question1 is _ and
   check( question1, =, peaceCorps ) and
   Conclusion = 'No_Payment_Due'  .

relation q_question1( Conclusion ) if
   the answer to question1 is _ and
   check( question1, =, disability ) and
   Conclusion = 'No_Payment_Due'  .

relation q_question1( Conclusion ) if
   the answer to question1 is _ and
   check( question1, =, student ) and
   Conclusion = 'No_Payment_Due'  .

relation q_question1( Conclusion ) if
   the answer to question1 is _ and
   check( question1, =, financial ) and
   Conclusion = 'No_Payment_Due'  .

relation q_question1( Conclusion ) if
   the answer to question1 is _ and
   check( question1, =, military ) and
   Conclusion = 'No_Payment_Due'  .

relation q_question2( Conclusion ) if
   the answer to question2 is _ and
   check( question2, =, notContinuous ) and
   Conclusion = 'Payment_Due'  .

relation q_question2( Conclusion ) if
   the answer to question2 is _ and
   check( question2, =, continuous ) and
   Conclusion = 'No_Payment_Due'  .

group group1
   notContinuous, continuous .

question question2
   'ContinuouselyEnrolled?~M~J' ;
   choose one of group1
   because 'ContinuouselyEnrolled' .

group group2
   na, peaceCorps, financial, military, student,
disability .

question question1
   'Deferment Type?' ;
   choose one of group2
   because 'Deferment Type' .
```

Fig. 6. A Prolog code generated from the VisiRule to be used as the KBS for student loan payment decision.

D. *Knowledge Consultation Phase*

Once a Prolog code for the KBS is automatically created, the system is ready for users to consult the student loan payment decision. The GUI of the knowledge consultation is illustrated in Fig. 7.

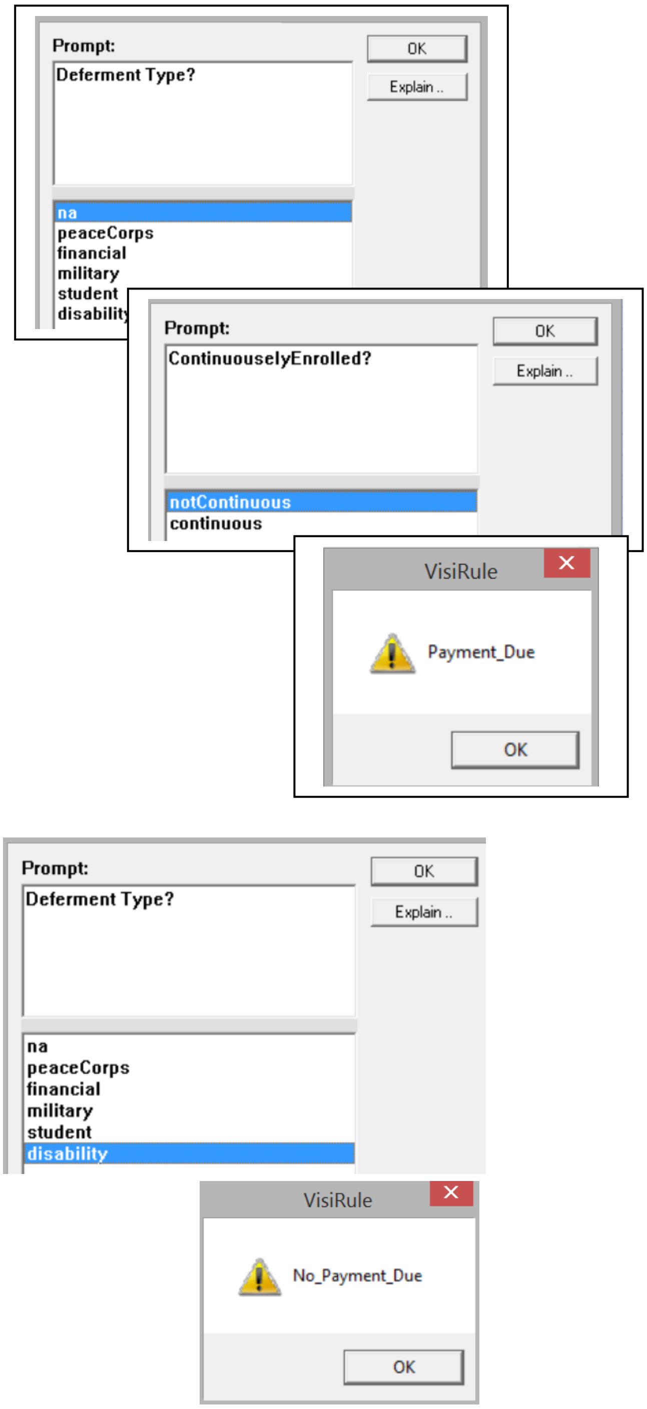


Fig. 7. Example of KBS usage to consult the student loan payment decision.

V. Conclusion

In this research work, we explain the design of a knowledge based system (KBS) that can be automatically created from the data records. The key components of our automated KBS construction are the combination of data mining technology and a self-generated logic program from the visualized tool. We choose a logic programming paradigm because logic is a fundamental aspect of most knowledge intensive applications. If we can induce a model and represent it in a format appropriate for knowledge transferring, an intelligent tool such as WIN-PROLOG can be applied to generate a shell for the KBS.

We demonstrate this idea through the case study of student loan payment decision. The visual tools such as KNIME and VisiRule of WIN-PROLOG are proven suitably fit our design of the automated KBS construction. At present our knowledge transferring module is semi-automatic. We plan to automate this process in our future research.

Acknowledgment

The research work has been funded by grants from the National Research Council of Thailand (NRCT) and Suranaree University of Technology. The first author has been supported through the funding of Knowledge Engineering Research Unit. The second author has been supported by the Data Engineering Research Unit.

References

[1] M. Akram, I.A. Rahman, and I. Memon, “A review on expert system and its applications in civil engineering,” Int J of Civil Engineering and Built Environment, vol.1, no.1, 2014, pp. 24-29.

[2] F. Alonso, L. Martinez, A. Perez, and J.P. Valente, “Cooperation between expert knowledge and data mining discovered knowledge: lesson learned,” Expert Systems with Applications, vol.39, 2012, pp. 7524-7535.

[3] A.B. Badine, Expert Systems: Applications in Engineering and manufacturing, Prentice Hall, 1992.

[4] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kotter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, “KNIME – the Konstanz information miner: version 2.0 and beyond,” ACM SIGKDD Explorations Newsletter, vol.11, no.1, 2009, pp. 26-31.

[5] Y. Chen, H. Lai, and H. Lin, “Constructing product knowledge-sharing system for Internet transaction-matching model,” J of Convergence Information Technology, vol.5, no.9, 2010, pp. 135-145.

[6] I. Chorbev, D. Mihajlov, and I. Jolevski, “Web based medical expert system with a self training heuristic rule induction algorithm,” Proc. 1st Int Conf on Advances in databases, Knowledge, and Data Applications, 2009, pp. 143-148.

[7] P. Clark, N. Balasubramanian, S. Bhakthavatsalam, K. Humphreys, J. Kinkeal, A. Sabharwal, and O. Tafjord, “Automatic construction of interface-supporting knowledge bases,” Proc. 4th Workshop on Automated Knowledge Base Construction, 2014.

[8] B.R. Gaines and M.L.G. Shaw, “Eliciting knowledge and transferring it effectively to a knowledge-based system,” IEEE Transactions on Knowledge and data Engineering, vol.5, no.1, 1993, pp. 4-14.

[9] G. Holmes and S.J. Cunningham, “Using data mining to support the construction and maintenance of expert systems,” Proc. 1st New Zealand

International Two-Stream Conference on Artificial Neural Networks and Expert Systems, 1993, pp. 156-159.

[10] X. Huang and J.R. Jensen, “A machinelearning approach to automated knowledge-base building for remote sensing image analysis with GIS data,” Photogrammetric Engineering & Remote Sensing, vol.63, no.10, 1997, pp. 1185-1194.

[11] KNIME Analytics Platform [https://www.knime.org/knime].

[12] C. Leon, F. Biscarri, I. Monedero, J.I. Guerrero, J. Biscarri, and R. Millan, “Integrated expert system applied to the analysis of non-technical losses in power utilities,” Expert Systems with Applications, vol.38, 2011, pp. 10274-10285.

[13] M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California.

[14] R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum, and J. Lederberg, “DENDRAL: a case study of the first expert system for scientific hypothesis formation,” Artificial Intelligence, vol.61, no.2, 1993, pp. 209-261.

[15] Logic Programming Associated Ltd, WIN-PROLOG 5.0 [http://www.lpa.co.uk/win.htm].

[16] J. Mayfield, P. McNamee, C. Harman, T. Finin, and D. Lawrie, “KELVIN: extracting knowledge from large text collections,” Proc. 2014 AAAI Fall Symposium on Natural Language Access to Big data, 2014, pp. 555-570.

[17] A.H. Mohammad and N.A.M. Al Saiyd, “A framework for expert knowledge acquisition,” Int J of Computer Science and Network Security, vol.10, no.11, 2010, pp. 145-151.

[18] M.J. Pazzani and C.A. Brunk, “Detecting and correcting errors in rule-based expert systems: an integration of empirical and explanation-based learning,” Knowledge Acquisition, vol.3, 1991, pp. 157-173.

[19] S. Riccucci, A. Carbonaro, and G. Casadei, “Knowledge acquisition in intelligent tutoring system: a data mining approach,” Proc. 12th Int Conf on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, 1999, pp. 1195-1205.

[20] W.R. Swartout, “Future directions in knowledge based systems,” ACM Computing Surveys, vol.28, no.4, article 13, 1996.

[21] D.Z. Wang, Y. Chen, S. Goldberg, C. Grant, and K. Li, “Automatic knowledge base construction using probabilistic extraction, deductive reasoning, and human feedback,” Proc. Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, 2012, pp. 106-110.

[22] T. Witkowski, P. Antczak, and A. Antczak, “Machine learning-based classification in manufacturing system,” Proc. 6th IEEE Int Conf on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2011, pp. 580-585.

[23] Y. Yoon, R.W. Brobst, P.R. Bergstresser, and L.L. Peterson, “Automatic generation of a knowledge-base for a dermatology expert system,” Proc. 3rd Annual Symposium on Computer-Based Medical Systems, 1990, pp. 306-312.

[24] A. Zhu, J. Li, and T. Leong, “Automated knowledge extraction for decision model construction: a data mining approach,” Proc. AMIA Annual Symposium, 2003, pp. 758-762.