

Data Mining: Definite Rules Generation Using a Rough Sets Approach

JAMILA RUTH A. HOJAS

Abstract


In this study, a program is created to extract definite rules based on a rough sets framework in the context of data-mining.

Introduction

Rapid advances in database technology have given way to vast databases which usual methods of data analysis can no longer handle. Thus, a new challenge for scientists today is to seek a method for the automation of data analysis. Modern developments in this area have led to a field called data-mining.

Data-mining is a process which applies specific algorithms to fit models to data or to discover interesting and possibly useful patterns that could provide information about relationships between the data. It is one of the steps in a larger process called KDD (Knowledge Discovery in Databases) and combines methods taken from areas such as statistics and artificial intelligence. One of these methods is the rough set theory.

Rough Sets is a set theory that classifies objects into sets based on attributes of the objects. It represents a new mathematical approach to vague-

 JAMILA RUTH A. HOJAS (jam@sun.msuiit.edu.ph), faculty, Computer Science Department, College of Engineering, MSU-Iligan Institute of Technology, has an M.S. in Computer Science (2000) from the Ateneo de Manila University, Quezon City, Philippines.

ness and uncertainty. The theory was originated by Zdislaw Pawlak in 1970's as a result of a long term program of fundamental research on logical properties of information systems. The methodology is concerned with the classificatory analysis of imprecise, uncertain or incomplete information or knowledge expressed in terms of data acquired from experience. Primary notions of the theory are the approximation space and lower and upper approximations of a set.

It was this theory that formed the basis for Torulf Mollestad and Andrzej Skowron's formal framework [16] that specifies the automated transformation of data into knowledge. The framework shows how definite and default rules can be generated for any database by examination of its contents. It uses a data reduction technique to eliminate unnecessary information to make the rules as short and concise as possible. Definite rules provide a view of the consistent knowledge found in a database while default rules (also known as "indefinite rules") determine the accuracy of imprecise information in a database. Thus, it defines how to distinguish (quantitatively) what information can be considered and cannot be considered as knowledge.

In this study, a data-mining program is created to extract definite rules based on the framework of Mollestad and Skowron [16].

Main Objective

The main objective of this study is to model the deterministic data of a database by creating a program that extracts definite rules from it using a rough set framework.

Scope and Limitations

The framework for the data mining program is based on the Rough Sets theory. Although the Rough Sets theory is ideal for noisy data sets and generating default rules, only definite rules are generated in this case. If changes are made to the data in the database or new data are added, definite rules will have to be generated again. The program itself is not equipped with a GUI (Graphical User Interface). Furthermore, the definite rules generated are not in the form of a running program that can be tested automatically on test data but must be converted to a program in order to test their effectiveness.

Literature Search

Data-mining is used to refer to the step in a larger process called KD (Knowledge Discovery) in which specific algorithms are applied to fit models to data or to discover interesting and possibly useful patterns that could provide information about relationships between the data. The basics of data mining are techniques from the fields of Machine Learning and Statistics [21]. A wide range of data mining tools – such as neural networks, rule-based systems, decision trees, genetic algorithms, statistical applications, theory of rough sets – alone or in combination, may be applied to a problem [21].

GA-MINER [15] uses a genetic algorithm-based approach to data mining. In [14] Srikant, Vu, and Agrawal use the Apriori Algorithm to mine association rules. They then compared its integration with the Reorder Algorithm versus its integration with algorithm MultipleJoins to place item constraints on the mined association rules. Heckerman [12] shows that Bayesian networks are also suitable for modeling relationships between data. In a paper by Glymour, Madigan, and Pregibon [11], statistical themes and lessons that are directly relevant to data mining are identified. The Advanced Scout application (AS) uses a technique called Attribute Focusing (AF) [6] that compares the overall distribution of an attribute with the distribution of this attribute for various subsets of the data. If a certain subset of data has a characteristically different distribution for the focus attribute, then that combination of attributes is marked as interesting. In this project, a Rough Set approach will be used for data mining definite rules. Ideas were taken from different applications that used the rough set as the underlying principle. In [19] Tanaka and Tsumoto describe the relations between rough set theory and rule-based description of neurological diseases. A study by Patterson [22] used the original model of rough sets for data analysis of objective clinical findings from pneumonia patients. Mollestad and Skowron [16] provide a framework for definite and default rule extraction using the rough set theory which is used as a framework for the definite rule extraction program that is created in this project. Additional papers by Pawlak [13, 17, 18] provide the theoretical foundation of the rough set theory.

In [21], different rough sets applications are described. RSES which is available for Hewlett Packard work stations can process up to 16,000 attributes with a maximum of 30,000 objects. DataLogic/R made by Reduct

Systems Inc. was written in C. It takes techniques from knowledge representation, inductive logic and rough sets. For its PC version, it can process up to 2000 attributes. The KDD-R is a system implemented under UNIX. It is based on the Variable Precision Rough Set (VPRS) model. LERS (Learning from Examples on Rough Sets) induces rules. It calculates the lower and the upper approximation and generates the deterministic and indeterministic rules.

Rule Generation

To summarize, the process for rule generation basically starts from the selection of the condition and decision attributes which partitions the rows in the database into what's termed as equivalence classes. The reason for doing so is because, ultimately, the goal will be to discover rules that will predict the values of decision attributes given the values of condition attributes. The next step is the data reduction process which reduces the number of conditions by eliminating the non-essential condition attributes. This is done using the discernibility matrix and boolean algebra. From Mollestad and Skowron's framework, a function called the rough membership function is used to determine whether an equivalence class is deterministic or not. It is deterministic if the rough membership function is one, if it is in between zero and one (1), then, the equivalence class is indeterministic. Definite (to model the deterministic data), indefinite (to model the indeterministic data), and default rules are then generated. In this study, only definite rules are generated.

Programming Modules

The program modules are created using the C language in a UNIX environment to maximize memory allocation and usage.

Diagram 1 below illustrates the steps and programming modules involved in rule generation. The extraction of definite rules from a database occurs in four (4) stages. In the first stage, equivalence classes are generated using an RDBMS (Relational Database Management System). The results are saved to a text file in a standard delimited format. In stage two (2), the 'genmat' executable module processes the text file from stage one and generates a discernibility matrix which is saved to a file. The goal of stage three (3) is to generate the relative discernibility functions from the discernibility

matrix. The executable program 'genrdf' performs this. The relative discernibility functions are stored to a file with a '.rdf' extension. Finally, in stage four (4), from these relative discernibility functions, executable program 'genrules' generates the definite rules. The rules are saved to a file and have an IF-THEN format like that of rules in VP-Expert [25].

Example

Figure B.9 shows what the contents of a '.rules' file of stage four (4) might look like. In this example, six definite rules were generated from an 8124 row database. The rules define a mushroom's 'CLASS' (edible or poisonous) based on two properties: CAP-SHAPE and CAP-SURFACE. Notice that some rules are redundant (Rule 1, Rule 3, and Rule 4 for example).

Results and Discussion

Based on the given rough set framework a process for definite rules generation was identified and implemented. The created C program modules are able to derive definite rules of a database given a set of generated equivalence classes using the RDBMS software MS Access. The type of data that can be used for the above procedure though is limited to data for classification problems and data that is not continuous-real valued. Rules generated for data other than that mentioned above would not make much sense.

The number of condition and decision attributes is severely limited since the representation of the discernibility matrix is in text form. That is, it is represented as a text file consisting of a string of 0's and 1's therefore extremely limiting the size of the associated .mat file and taking a considerable amount of time to process as the .mat file grows. So far this researcher was able to process at most fifteen (15) decision attributes versus one condition attribute due also to the limitations of the hardware being used.

The rules that are generated are sometimes redundant. That is, a rule generated for an equivalence class is exactly the same rule generated for a different equivalence class. This is due to the process of the data or attribute reduction technique used in the framework. In this situation it is the case wherein the only attribute whose value makes two equivalence classes different is eliminated thus resulting in redundant rules.

Overall, though the process of definite rule generation is not automatic,

the researcher has been able to create program modules and identify existing software that may be utilized as tools in the data-mining process.

Recommendations

First of all, a module to eliminate the redundant rules can be created. Second, representing the string of 0's and 1's as bits instead of characters will vastly reduce hard disk space usage and run time during processing. Third, the programs can be extended to include the generation of indefinite rules. Fourth, a module to test the rules generated is also recommended.

Acknowledgement

I wish to acknowledge my adviser Dr. Pablo R. Manalastas of Ateneo de Manila University for his guidance throughout this study. Also with heartfelt gratitude, I would like to thank Mr. Torulf Mollestad and Mr. Andrzej Skowron for kindly sharing their paper on their rough set framework to everyone on the internet. Thank you for your generosity! Thank you to MSU-IIT's College of Engineering for the yearly forum that enables its faculty to present their studies and the initiative of the organizers to publish the proceedings. Thanks to Dr. Emmanuel Lagare and Prof. Eli Mostrales for their helpful comments and corrections to this paper. Finally, I thank my husband Michael for his wonderful support.

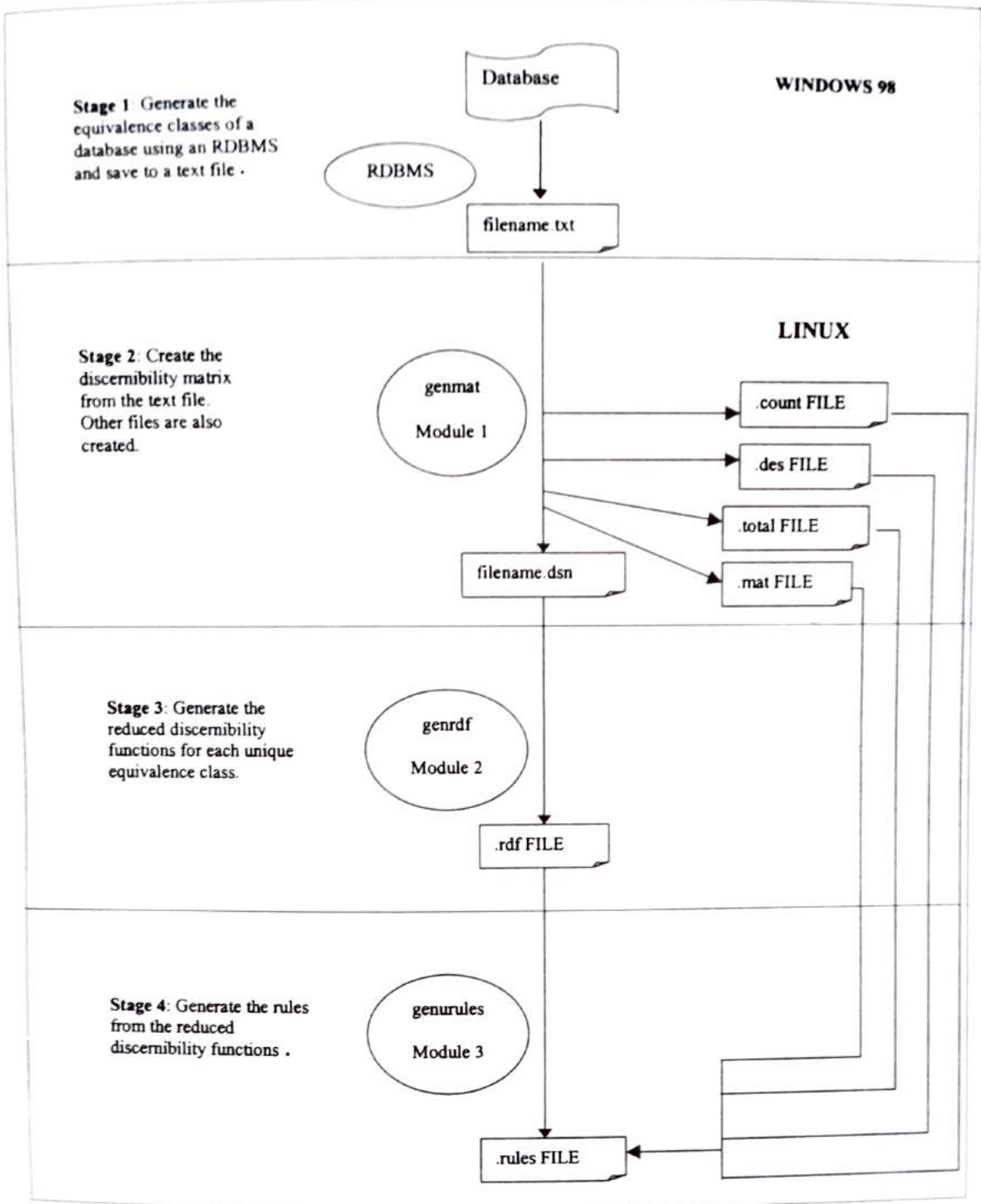


Diagram 1. Definite Rules Generation Process Flow Chart

```

DEFINITE RULES GENERATED FROM: surf
TOTAL ROWS: 8124
CONDITION ATTRIBUTES: 2

RULE 0 (EQUIVALENCE CLASS 5 : FREQUENCY 32/8124 )
  IF      CAP - SHAPE = sunken
  THEN    CLASS = edible;

RULE 1 (EQUIVALENCE CLASS 6 : FREQUENCY 1/8124 )
  IF      CAP - SURFACE =      grooves
  THEN    CLASS = poisonous;

RULE 2 (EQUIVALENCE CLASS 7 : FREQUENCY 1/8124 )
  IF      CAP - SURFACE = grooves OR
          CAP - SHAPE = conical
  THEN    CLASS = poisonous;

RULE 3 (EQUIVALENCE CLASS 8 : FREQUENCY 1/8124 )
  IF      CAP - SURFACE = grooves
  THEN    CLASS      = poisonous;

RULE 4 (EQUIVALENCE CLASS 9 : FREQUENCY 1/8124 )
  IF      CAP - SURFACE = grooves
  THEN    CLASS = poisonous;

RULE 5 (EQUIVALENCE CLASS 11 : FREQUENCY 3/8124 )
  IF      CAP - SHAPE = conical
  THEN    CLASS = poisonous;

```

Figure B.9. SURF-SHP.RULES

References

Books:

- [1] Dean, Thomas (1995), et. al. *Artificial Intelligence: Theory and Practice*, California: Addison-Wesley, USA.
- [2] Kimball, Ralph (1996), *The Data Warehouse Toolkit: Practical Techniques For Building Dimensional Data Warehouses*, New York: John Wiley & Sons, Inc.
- [3] Microsoft Corporation (1994), *Microsoft Access: User's Guide*, USA.
- [4] Tanenbaum, Andrew S. (1990), *Structured Computer Organization*, 3rd. Ed., Philippines: Prentice-Hall International.
- [5] Qualline, Steve (1995), *Practical C++ Programming*, California: O'Reilly & Associates, Inc.

Journals

- [6] Bhandari, I., Colet, E., Parker, J., et. al.. (1997), Advanced Scout: Data Mining and Knowledge Discovery in NBA Data, *Data Mining and Knowledge Discovery*, Vol. 1, pp. 121-125.
- [7] Edelstein, Herb, & Millenson, J. (1997), Lessons from the Trenches: Knowledge, Discovery, and Data Mining, *DBMS*, February.
- [8] Edelstein, Herb (1996), Data Mining: Exploiting the Hidden Trends in Your Data., *DB2 Online Magazine*, Spring Issue. (<http://www.db2mag.com/9701edel.htm>)
- [9] Fayyad, Usama M. (1997), Editorial, *Data Mining and Knowledge Discovery*, pp.1, 5-10.
- [10] Fayyad, Usama M. (1997), Editorial, *Data Mining and Knowledge Discovery*, Vol. 1, Issue 3.
- [11] Glymour, C., Madigan, D., Pregibon, D., et.al. (1997), tatistical Themes and Lessons for Data Mining, *Data Mining and Knowledge Discovery*, pp.1, 11-28.
- [12] Heckerman, David (1995), Bayesian Networks for Data Mining, *Data Mining and Knowledge Discovery*, pp.1, 79-119.
- [13] Pawlak, Zdislaw (1995), Vagueness and Uncertainty: A Rough Set Perspective., *Computational Intelligence*, Vol. 11, pp. 227-232.

Papers

- [14] Agrawal, R., Srikant, R., Vu, Q. (1997), Mining Association Rules with Item Constraints, *Proceedings of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining*, Newport Beach, California, August.
- [15] Flockhart, Ian W., & Radcliffe, Nicholas J., A Genetic Algorithm-Based Approach to Data Mining, Department of Mathematics and Statistics, University of Edinburgh, UK. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, 1996. At: <http://www.informatik.uni-trier.de/~ley/db/conf/kdd/FlockhartR96>.
- [16] Mollestad, Torulf, & Skowron, Andrzej., A Rough Set Framework for Data Mining of Propositional Default Rules. *Foundations of Intelligent Systems, 9th International Symposium, ISMIS '96, Zakopane, Poland, June 9-13, 1996, Proceedings. Lecture Notes in*

- Computer Science*, Vol. 1079, Springer, 1996. At: <http://www.acm.org/sigs/sigmod/dblp/db/conf/ismis/ismis96.html>.
- [17] Pawlak, Zdislaw., A Rough Set Approach to Knowledge-Based Decision Support, Institute of Computer Science, Warsaw University of Technology. (ftp://ftp.ii.pw.edu.pl/pub/Reports/10_95_.ps.Z)
- [18] Pawlak, Zdislaw., Rough Sets, Rough Relations and Rough Functions, Institute of Computer Science, Warsaw University of Technology and Institute of Theoretical and Applied Informatics Polish Academy of Sciences. (ftp://ftp.ii.pw.edu.pl/pub/Reports/24_94.ps.Z)
- [19] Tanaka, H., & Tsumoto, S., Induction of Disease Description based on Rough Sets, Department of Information Medicine, Medical Research Institute, Tokyo Medical and Dental University.

Web Sites

- [20] *Advances in Knowledge Discovery and Data Mining*. <http://www.aai.org/Press/Books/Fayyad/fayyad.html>. Accessed 24 May 1998.
- [21] Grossman, Robert, et al. *Data Mining Research: Opportunities and Challenges. A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data*. <http://www.ncdm.uic.edu/m2d-finalreport.html>. Accessed January 1999
- [22] Paterson, Grace I. *A Rough Sets Approach to Patient Classification in Medical Records*. . <http://www.mcms.dal.ca/dme/mi95pat1.html>
- [23] Solheim, Helge Grenager. *Data Mining Process*. <http://www.pvv.unit.no/~hgs/node80.html> . Accessed February 1998
- [24] *A Brief Introduction to Rough Sets*. Electronic Bulletin of the Rough Set Community EBRSC 1993. Accessed February 1998.
- [25] *A VP-Expert Primer*. http://www.cis.ysu.edu/~john_824/vpxguide.html. Accessed November 2001.