

Using rough set as a tool for knowledge discovery in DSS

Ahmad Farhan Bin Jaafar¹, Jamilin Jais¹, Mohd Hakim Bin Haji Abdul Hamid¹, Zarina Binti Abdul Rahman¹, Djamel Benaouda, Dr.¹

¹ College of Information Technology, Universiti Tenaga Nasional, Km 7, Jalan Kajang-Puchong, 43009 Kajang, Selangor, Malaysia, paan@uniten.edu.my

Data mining and/or knowledge discovery is a very hot issue nowadays, as more and more information is being stored digitally the ability to collect data far outweighs the ability for a person to analyze it. Therefore data mining and knowledge discovery is a very important part of today's business. The purpose of this is to find descriptive and predictive model and/or pattern from the data. Descriptive pattern are used when trying to classify a new data that will be introduced into the system and predictive pattern are used to forecast possible future outcome.

Rough set theory was developed by Zdzislaw Pawlak in the early 1980's. Rough set deals with classification of discrete data table in a supervised learning environment. Although in theory rough set deals with discrete data, rough set is commonly used in conjunction with other technique to do discretization on the dataset. The main feature of rough set data analysis is non-invasive, and the ability to handle qualitative data. This fits into most real life application nicely.

This paper will go through the process of applying data mining technique using rough set on a TNB (the local electricity company) fault reporting data. The aim of the project is provide a DSS system in which the patterns of the faults in relation to time and/or place can help executives of the company to improve customer service. Patterns in the data can help in identifying problematic geographic locations so that infrastructure in those locations can be upgraded, or help identify peak usage hours when faults are common. With these information it will be easier to find out in which areas do the executives need to improve on to provide better service to the customers.

Keywords: Data mining, knowledge discovery in database (KDD), rough set, DSS, CRM

1. INTRODUCTION

Rough set theory was developed by Zdzislaw Pawlak in the early 1980's. Rough set deals with classification of discrete data table in a supervised learning environment. Although in theory rough set deals with discrete data, rough set is commonly used in conjunction with other technique to do discretization on the dataset. The main feature of rough set data analysis is non-invasive, and the ability to handle qualitative data. This fits into most real life application nicely. Rough set have seen light in many researches but seldom found its way into real world application.

Knowledge discovery with rough set is a multi-phase process consisted of mainly:

- Discretization
- Reducts and rules generation on training set
- Classification on test set

The data that will be used is the research is from TNB's customer power outage complain form. The purpose of this is to see weather there is any geographical or chronological correlation with power outage. If pattern emerges then this could help TNB in determining problem areas in terms of physical line problems or in determining when certain time of day when outages are prone to happening. This infor-

mation can be helpful in improving quality of service to the customer. The data itself is very large, containing over 400,000 outage records from TNB from the past year.

2. THE DATASET

The dataset used for the project is obtained from the TNB's fault reporting database. The original dataset consist of over 400,000 records. The records cover all fault report from all over Malaysia.

2.1 Cleaning of data

The original dataset contained fault reports from all over Malaysia, which amounts to over 400,000 records. The sheer size of the database makes manipulating it a problem. So it is decided that only a subset of the data is used in the data mining operation. The subset of the dataset that will be used for the data mining operation will be chosen from all records from Selangor. This cut down the dataset to a manageable size of around 80,000 records.

The dataset consists of 60 columns in total. A lot of these columns are either superfluous, or completely useless for the purpose of data mining. The main types of data that needed to be cleaned out are

- Empty columns
No data exist for any entry in the dataset
- Almost empty columns
These columns are not empty but might as well been, some columns have 80k++ empty entries out of 86,360 entries, these falls under mostly empty columns. For these columns, I've pick 80k as my cut off point for column to fall under mostly empty to be removed.
- Memo columns
Columns that are used to describe to the users about something example: "Turn right at Jalan Reko", "customer called twice", etc. etc.
- Misc columns
Misc data are not really useless, but they provide little use in data mining. Personal information like phone number, address (dataset keeps track of the zone and district of the caller, so address is not needed to determine location), and name of the caller. Database generated guid and or reference number, e.g. complaint number, is useless for data mining. And finally the state is also useless because the data is already filtered to only Selangor state

3.0 MINING PROCESS

The process of KDD using rough set is a multi step process that require a few steps.

3.1 Discretization

Because rough set theory is a symbolical method rather than a numerical method, roughest theory cannot process continuous data. Discretization is a process that converts continuous data into discreet intervals to be used in roughest. There a couple of popular techniques that is used to discretize data. The project will use Boolean reasoning technique to do the discretization on the data. The technique is moderately simple with good results seen on most dataset. [6].

"The tests are showing that they are very efficient from the point of view of time complexity."[6]
"The heuristics for symbolic value partition allow to obtain more compressed for of decision algorithm. Hence, from the minimum description length principle, one can expect that they will return decision algorithms with high quality of unseen object classification."[6]

3.2. Reducts and rules generation

The reducts and rules generation is the core of the roughset. In this part, the algorithm will go through the dataset to generate reducts and rules. In this project the algorithm used will be the Holte's 1r algorithm. The algorithm is a simple and fast and provides compatible result when compared to other more advanced technique. Past research [4] showed that despite being a simple algorithm, the Holte's 1r algorithm is good enough to produce very good results.

```

For each attribute a, form a rule as follows:
  For each value v from the domain of a,
    Let c be the most frequent class in the set of instances where a has value v.
    Add the following clause to the rule for a:
      If a has value v then the class is c
    Calculate the classification accuracy of this rule.
Use the rule with the highest accuracy

```

Figure 1: Pseudocodes for the Holte's 1r Reducer [4]

4. DESIRED OUTCOME

The desired outcome of the system will mostly be information that can help the TNB executive in assigning repair team schedule. This include information like the correlation between response time and type of fault, time of day, location of fault etc. etc. Other than that, the information needed also include information that can help executives identify certain geological (location) or chronological (time) patterns in fault reporting. This will help them to identify problems areas in terms of frequent outage etc. etc. This will help them to improve quality of service to the customer.

The desired outcome for the rules generation process is, but not limited to:

- Correlations between
 - Location-> Time Team Arrive
 - Problem type ->Time needed to complete work
 - Time of day ->Type of problem
 - Location -> Time of problem
 - Location-> Number of problems
- Forecast of number of problem on a given timeframe (every month, during the morning)

5. VISUALIZATIONS

The project presents data in the form of charts and graphs. For frequency data on different faults type on specific location. The project will use pie chart as a form of visualization. For fault frequency based on location, the project will present the user with a map of the area with different sized colored circles to show frequency of fault happening on certain locations. Overall the system try to provide multiple views to the user based on selected item and give ability to customize the view according to the data that the user wish to see.

6. System architecture

The overview of the architecture of the system can be seen in figure. The proposed architecture will adopt the traditional architecture of a data mining system. Data from multiple channels is collected on the operational data store for fast transaction and up to date data that can be used for the front office. Then, periodically, the data is extracted, cleans, transformed and imported into the data warehouse. The data will then will be send to the appropriate data marts for departmental use. Then, according to the needs of the user, either the enterprise data or the departmental data is sent to the OLAP tier for process-

ing. The results is then stored and then sent to the decision makers through the use of thin clients. The overview of this architecture is seen in Figure 2 (And Appendix A). The proposed system is pretty good in theory as it provides compartmentalization of data and collection of data from multiple channels. The architecture is simple and sticks to the basis of founded work [1], [8], [9], and should provide a good base for the system.

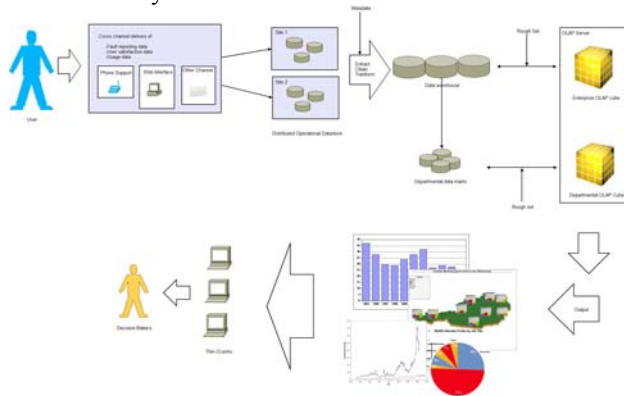


Figure 2: The overview of the proposed system architecture

7. CONCLUSIONS

The project shows that roughset theory can be used as a tool for knowledge discovery. Even though it is a symbolical method, application of a suitable quantization technique will allow it to perform on just about any type of data. As opposed to numerical method that cannot be adapted to be used for symbolical data. Roughset provide a useful tool that can be used on a lot of different data regardless weather it is numerical or symbolical and it also provide a non-intrusive methodology to knowledge discovery.

REFERENCES

- [1] Berry, Michael J. A (1997) Data mining techniques: for marketing, Sales and customer support, New York: John Wiley
- [2] I. Düntsch. G. Gediga (1999) Rough set data analysis: A road to non-invasive knowledge discovery
- [3] Holmes, G., and Nevill-Manning, C.G. (1995). Feature Selection via The Discovery of Simple Classification Rules. Proc. International Symposium on Intelligent Data Analysis (IDA-95), Baden-Baden, Germany.
- [4] Holte R.C (Machine Learning., vol. 11, pp. 63--91, 1993) Very simple classification rules perform well on most commonly used datasets
- [5] H. S. Nguyen and A. Skowron. Quantization of real-valued attributes. In Proc. Second International Joint Conference on Information Sciences, pages 34–37, Wrightsville Beach, NC, Sept. 1995.
- [6] H.S. Nguyen and A. Skowron. *Boolean reasoning for feature extraction problems*. 10th International Symposium on Methodologies for Intelligent Systems (ISMIS'97), volume 1325 of Lecture Notes in Artificial Intelligence, pages 117--126, Berlin, 1997
- [7] Aleksander Øhrn. ROSETTA Technical Reference Manual. Knowledge Systems
- [8] Berry, Michael J. A. (2000) Mastering data mining: the art and science of customer relationship management, New York: John Wiley
- [9] Marakas, George M. (1999) Decision support systems in the twenty-first century: DSS and data mining technologies for tomorrow's manager, New Jersey: Prentice Hall

Appendix A The overview of the proposed system architecture

