Exploration Mining in Diabetic Patients Databases: Findings and Conclusions

Wynne Hsu, Mong Li Lee, Bing Liu and Tok Wang Ling

School of Computing
National University of Singapore
Lower Kent Ridge Road, Singapore 119260

{whsu, leeml, liub, lingtw}@comp.nus.edu.sg

Abstract

Real-life data mining applications are interesting because they often present a different set of problems for data miners. One such real-life application that we have done is on the diabetic patients databases. Valuable lessons are learnt from this application. In particular, we discover that the often neglected pre-processing and post-processing steps in knowledge discovery are the most critical elements in determining the success of a real-life data mining application. In this paper, we shall discuss how we carry out knowledge discovery on this diabetic patient database, the interesting issues that have surfaced, as well as the lessons we have learnt from this application. We will describe a semi-automatic means for cleaning the diabetic patient database, and present a step-by-step approach to help the health doctors explore their data and to understand the discovered rules better.

1. Introduction

In Singapore, about 10 percent of the population is diabetic. This disease has many side effects such as higher risk of eye disease, higher risk of kidney failure, and other complications. However, early detection of the disease and proper care management can make a difference. To combat this disease, Singapore introduced a regular screening program for the diabetic patients in 1992. Patient information, clinical symptoms, eye-disease diagnosis and treatments are captured into a database. After eight years of data collection, a whole wealth of information has been gathered. This leads naturally to the application of knowledge discovery and data mining techniques to discover interesting patterns that exist in the data. The objective is to find rules that can be used by the medical doctors to improve their daily tasks, that is, to understand more about the diabetic disease or to find out something special about a particular patient population. Although knowledge discovery in databases has reported many successes in domains such as fraud detection, targeted marketing etc., we found that the application of data mining techniques to health sector has been relatively few in comparison. We believe this is primarily due to two reasons. First, the data captured by health clinics are typically very noisy. Many of the patient records contain typographical errors, missing values, or wrong information such as street names or date of birth etc; and worse, many records are in fact duplicate records. Cleaning these data takes tremendous amount of effort and time. In addition, many of the data collected are not in the forms that are suitable for data mining. They need to be transformed to more meaningful attributes before mining can proceed. Second, the health doctors are usually too busy seeing patients each day. They cannot afford the time or the energy to sieve through the thousands of rules generated by some state-of-the-art mining techniques on the diabetic patient database. Thus, it is important to present the discovered rules in some easy-to-understand fashion.

In this paper, we will demonstrate how we address these concerns. To overcome the problem of noisy data, we have developed a semi-automatic data cleaning system. The system reconciles database format differences by allowing the doctors to specify the mapping between attributes in different format styles and the encoding schemes used. Once the format differences have been reconciled, the problem of identifying and removing duplicate records is addressed. To resolve the problem of too many rules generated by the state-of-the-art mining techniques, we apply a user-oriented approach that provides step-by-step exploration of the data in order to better understand the discovered patterns.

2. Related Work

The medical domain offers a fertile ground for data mining applications. However, they have been few medical data mining applications as compared to other domains. [14] reported their experience in trying to automatically acquire medical knowledge from clinical databases. They did some experiments on three medical databases and the rules induced are used to compare against a set of pre-defined clinical rules. In the experiments, there is no direct involvement of the medical doctors. [11] focused on the needs to generate rules that does not violate the prior knowledge of the domain experts. The technique was applied to a Alzheimer's disease database. These applications focus on the generation of understandable rules.

While it is important to generate understandable rules, it is also important to the medical doctors to have a complete picture of all the rules that exist in the database. This leads naturally to association rule mining. The problem with association rules is, however, that there are often too many of them, and they also contain a large amount of redundancy [8].
Past research in dealing with this problem can be described with the following approaches:

(a) Discover all rules first and then allow the user to query and retrieve those he/she is interested in. The representative approach is that of templates [3]. This approach lets the user to specify what rules he/she is interested as templates. The system then uses the templates to retrieve the rules that match the templates from the set of discovered rules.

(b) Use constraints to constrain the mining process to generate only relevant rules. [12] proposes an algorithm that can take item constraints specified by the user in the association rule mining process so that only those rules that satisfy the user specified item constraints are generated. This also does not work well because doctors often do not have any specific rules to mine.

(c) Find unexpected rules. This approach first asks the user to specify his/her existing knowledge about the domain. The system finds those unexpected rules [5, 6, 13].

The first two approaches did not work well for our application because the doctors often do not know what they were looking for and thus could not give templates. The third approach works to a limited extend because the doctors are unable to specify too many existing beliefs. We noticed that the doctors are quite unwilling to specify any existing knowledge or requirement because it is a mental burden to them.

All the previous work in medical domain focus on the rule generation phase itself. In our experiences, we find that the process of reaching the stage where understandable rules can be generated and the process after the rules have been generated are, equally if not, more important than the rule generation stage itself. More emphasis needs to be given to the pre-processing stage and the post-processing stage.

3. Semi-Automated Data Cleaning

We embarked on the mining of diabetic patient database project in early 1999. We were given the diabetic patient database which contains about 200,000 screening records captured from 1992 - 1996. Each record has 60 fields. A preliminary analysis of the database reveals patient details such as their identification number[1], race, sex, date of birth, duration of diabetes, and the date of screening. Each screening involves an eye examination to detect signs of diabetes-associated eye diseases such as retinopathy, maculapathy and age-related eye diseases such as cataract and glycoma. The outcome of the eye examination is also recorded in the screening record.

Given the "garbage in, garbage out" principle, it is crucial for us to clean the diabetic patient database first before mining can proceed. A few observations were made in the cleaning process: (1) In real-life, due to the rapid pace of software and hardware upgrades, many database files undergo several upgrades within a short period of time with significant format

[1] Every Singaporean has a unique identity card number, which is similar to the social security number in the United States.

changes. For example, in our diabetic database, we encounter 4 different kinds of formats from 1992-1996; (2) With each format change, inconsistencies creep in such as dd/mm/yy or mm/dd/yy and use of abbreviations such as "ONE" vs "1"; (3) Many attribute fields are left blank, particularly in medical history related fields and treatment rendered fields, giving rise to a large number of missing values in the database.

To tackle these problems, we implemented a data cleaning system. This system allows a user to define mappings between attributes in different formats, the encoding schemes used, and whether the attributes are to be kept in the cleaned database. With this specification, a final standardized format schema is generated. Based on this standardized format schema, each of the database files are transformed accordingly into this standardized format. However, one major problem persists in the data after this standardization step – duplicate records. Having two or more records referring to a single screening session of a patient not only contributes to the problem of handling ever-increasing amount of data, but also leads to the mining of inconsistent or inaccurate information that is obviously undesirable. Note that a patient can have more than one screening at different dates. A Sorted Neighbourhood Method (SNM) is proposed to remove duplicate records [2]. The major steps are as follows:

Step 1. Choose one or more fields that uniquely identify each record in the database.

Step 2. Sort the database according to the chosen fields.

Step 3. Compare the chosen fields of the records within a sliding window. Possible duplicates are output to a file.

Step 4. User verifies the duplicates detected and true duplicates are removed from the database.

Based on the SNM method, additional techniques are employed to pre-process the data records before sorting them so as to increase the likelihood of the potentially duplicate records being brought to a close neighbourhood. These techniques include scrubbing data fields using external source files to remove typographical errors and the use of abbreviations, tokenizing data fields and then sorting the tokens in the data fields. For example, suppose Record 1 and Record 2 are duplicates containing the same data in all the fields except for the Name field which is {*Tan Lay-Hoon*} and {*Lay-Hoon Tan*} respectively. If the Name field is used to sort the database, then Record 1 and 2 will become very far apart; hence Record 1 and 2 will not be detected as duplicates. Sorting the tokens in the Name field will cause the Name field of Record 1 and 2 to become {*Lay-Hoon Tan*}. Utilizing these techniques in our data cleaning system, we must first choose an application-specific key to sort the database before the removal of duplicate records. This key is a sequence of a subset of attributes, or sub-strings within the attributes, which has sufficient discriminating power in identifying likely candidates for matching. There is no rule specifying how the key should be designed but it is important that users should choose the fields that contain representative information of the record to sort the database so that "potentially matching" records will be moved to within a close neighborhood.

In the diabetic patient database, we judiciously choose a set of attributes that will uniquely identify a patient's screening record. This key consists of the patient’s identification number (NRIC) and the date of screening (DOS). The NRIC number is unique to all Singaporeans or Singapore permanent residents. It is similar to the social security number of the United States of America. We pre-process these key fields with a reliable external source file containing the NRIC and name of persons. This is to remove any typographical errors in the NRIC or name fields in the database. The database is then sorted based on this key.

Next, pair wise comparison of nearby records is carried out. Figure 1 shows how a fixed-size sliding window is employed to limit the scope of record comparisons. Suppose the size of the window is *w* records, then every new record entering the window is compared with the previous *w-1* records to find "matching records", that is, records containing similar key values. The first record in the window then slides out of the window.

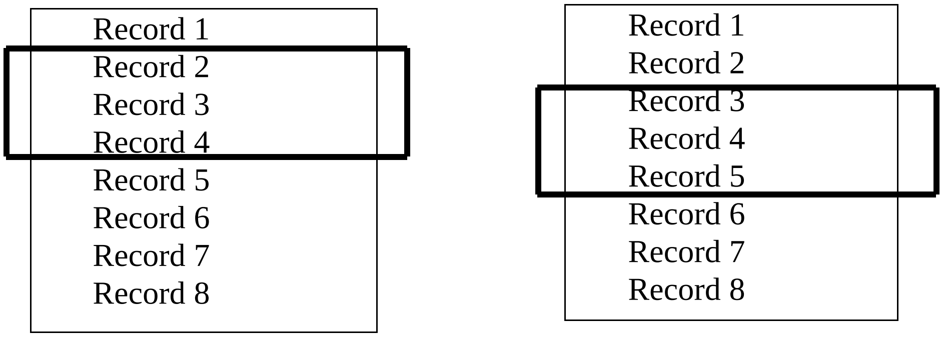


Figure 1: Record comparison with a sliding window of size 3

There are three types of duplicate records: true duplicates, false duplicates and duplicates that require further investigation. The attribute names and values have been encrypted to ensure confidentiality of data. The key used in this case is the combination of NRIC and DOS. Figure 2(a) shows a pair of true duplicates T1 and T2 where both the records contain the same data values in all the fields. Figure 2(b) illustrates the characteristics of false duplicates:

(i) Records F1 and F2 have the same values in fields such as the NRIC, date of birth (DOB), DOS, RACE and SEX which indicate that these two records refer to the same patient.

(ii) Fields such as CLIN, case type (C-TYPE), and screening results (HDUR, HVAUR, HVAUL, HVAPR, HVAPL) differ by one character.

(iii) The last field, REDR, which contains the reviewing doctor's last name, confirms that F1 and F2 are two separate screenings for the same patient.

The reason for such false duplicate detection is because NRIC and DOS combined together is not really a key. The key should be {NRIC, DOS, CLIN}. However, it is rarely the case that a person will go to two different clinics on the same day to do similar screening. For efficiency reasons, we choose {NRIC, DOS} as a key. Records such as F1 and F2 are the exceptional cases.

Finally, Figure 2(c) shows two records V1 and V2 that contain the same data in all the fields except for the screening result field HDUR. In this case, clarification from the expert is needed to find out whether the value of HDUR is 6 or 5.

Field	**Record T1**	**Record T2**
NRIC	t	t
RACE	C	C
CLIN	2	2
SEX	F	F
DOB	10/11/39	10/11/39
DOS	01/04/94	01/04/94
C-TYPE	1	1
C-SRC	G	G
HDUR	1	1
HVAUR	2	2
HVAUL	2	2
HVAPR	2	2
HVAPL	2	2
REDR	TEOH	TEOH

Figure 2(a): Record T1 and Record T2 are true duplicates

Field	**Record F1**	**Record F2**
NRIC	f	f
RACE	M	M
CLIN	1	2
SEX	M	M
DOB	10/11/33	10/11/33
DOS	01/04/94	01/04/94
C-TYPE	0	1
HDUR	1	3
HVAUR	2	3
HVAUL	2	3
HVAPR	2	3
HVAPL	2	3
REDR	TEOH	YEO

Figure 2(b): Record F1 and Record F2 are false duplicates

Field	**Record V1**	**Record V2**
NRIC	v	v
RACE	C	C
CLIN	2	2
SEX	F	F
DOB	11/27/30	11/27/30
DOS	10/01/93	10/01/93
C-TYPE	0	0
HDUR	6	5
:	:	:
:	:	:

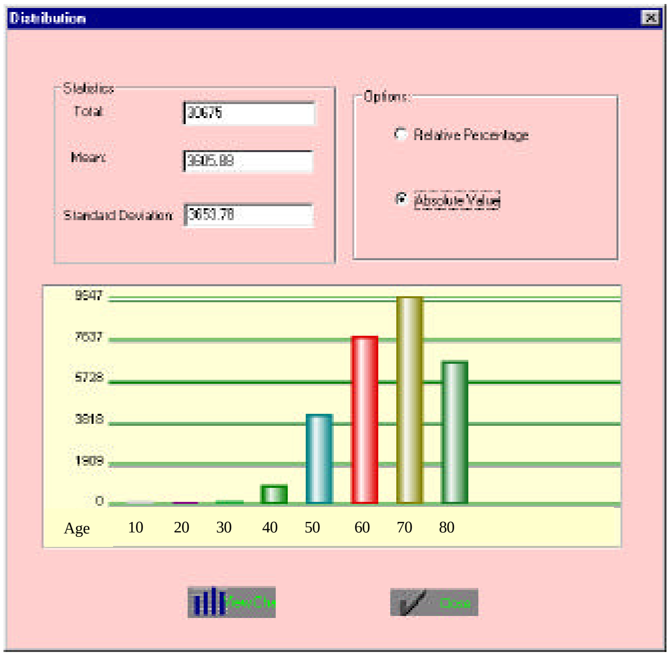
Figure 2(c): Verification needed to confirm if Record V1 and Record V2 are true duplicates

The experimental studies in [4] demonstrated that 70% of duplicate records are detected and removed as a result of the additional pre-processing steps. Note that in practice, it is inherently difficult to detect **all** the duplicates in a database.

4. Exploratory Mining of Diabetic Data

Once the data has been cleaned, the mining process began. Many data mining techniques are now available to discover patterns in data. In general, the discovered patterns fall into one of the following types: classification patterns, association patterns, sequential patterns, and spatial-temporal patterns. In the diabetic database application, we focus on the mining of classification and association patterns. Classification patterns provide a description of the characteristics of the population having certain diabetic-related eye disease. They are then used to predict whether a new patient is likely to have the eye disease. Association patterns provide a list of symptoms or treatments that often occur together, which give a complete picture of the relationships in the domain. A state-of-art data mining tool that integrates classification with association rule mining (CBA) is used to find all such patterns [7]. We used minimum support of 1% and minimum confidence of 50% as suggested by the doctors to mine association rules. Approximately 700 rules are generated in total. The doctors were totally overwhelmed. Some kind of post-processing is needed to help the doctors understand these rules. Our exploration mining methodology aims to give the doctors a better understanding of their data and the discovered patterns by helping the doctors to step through the massive amount of information in stages.

In the first stage, basic demographic information about the patients will be presented to the doctors. The goal here is to allow the doctors to have some basic idea about the impact of various single attributes, such as age, race, etc on the disease population. The information is presented in histogram graphs for easy digestion. Two types of information can be displayed: the absolute proportion (see Figure 3) or the relative proportion (see Figure 4). In Figure 4, we display the proportion of the different race-age populations having the eye disease. The race-age group with the highest disease



Distribution

Statistics

Total: 30675

Mean: 3605.89

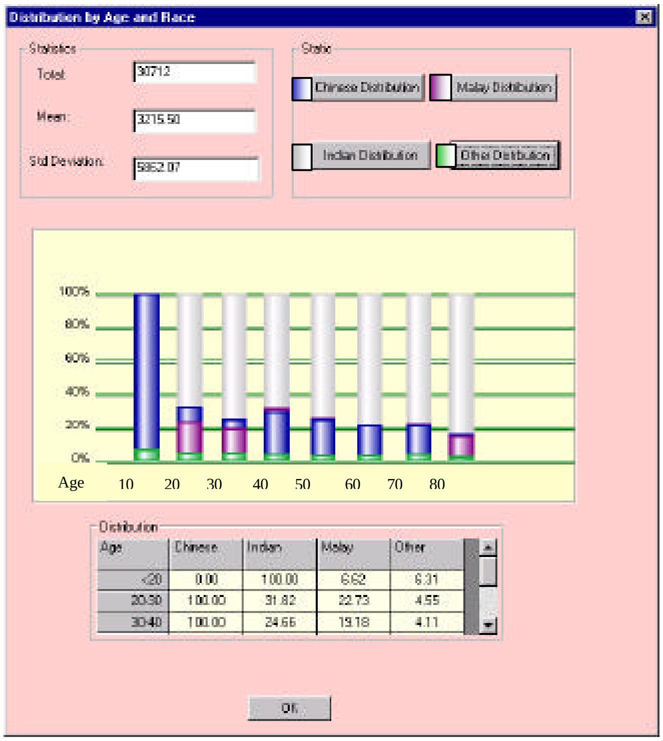
Standard Deviation: 3653.78

Options:

○ Relative Percentage

● Absolute Value

Figure 3. Absolute Proportion of Different Age Groups having Eye Disease.



Distribution by Age and Race

Statistics

Total: 30712

Mean: 3215.50

Std Deviation: 5852.07

Distribution

Age	Chinese	Indian	Malay	Other
<20	0.00	100.00	6.62	6.31
20-30	100.00	31.82	22.73	4.55
30-40	100.00	24.66	19.18	4.11

Figure 4. Relative Proportions of Different Races in Different Age Groups having Eye Disease.

proportion is assigned to be 100% while the rest of the group proportions are adjusted accordingly. Each bar represents an age group while the different colors within a bar denote different races. This provides a common basis for comparison. In one glance, the doctor is able to tell which factor dominates in terms of the number of patients having a particular eye disease and how much are the deviations among the groups over the years.

Once the doctors have obtained some basic idea of the demographic distributions, they are ready to investigate all significant correlations that exist in the data. This marks the beginning of phase two of the exploration mining. During this phase, the doctors are interested in knowing which set of symptoms often occur together and that if they do, it often implies the presence of some eye disease. Such correlations can be found using association rule mining. Association rule mining is commonly stated as follows [1]:

Let $I = \{i_1, \ldots, i_n\}$ be a set of *data fields*, and D be a set of data records. Each data record consists of a subset of attributes/fields in I. An *association rule* is an implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ holds in D with confidence c% if c% of data cases in D that support X also support Y. The rule has support s% in D if s% of the data case in D contains $X \cup Y$. The problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence.

Table 1 shows the breakdown of the number of association rules generated per year using the CBA mining tool [7] with a fixed class attribute, namely, the type of eye disease. The doctors are overwhelmed by the many rules and most of them do not make sense to the doctors. This is because these rules only indicate that there exist statistical relationships between the attributes, such as “the existence of A and B implies the existence of C”. However they do not specify whether the presence of A and B causes C, which is what the

Table 1. Number of rules generated.

Year	1992	1993	1994	1995	1996
Number of rules	109	168	169	181	160

doctors are interested in. Further analysis is needed to determine the potential factors of diabetic eye disease. This is the third phase of exploration mining. [15] studied how casual structures can be determined from association rules. A LCD algorithm is proposed for this purpose [15]. The LCD algorithm is a polynomial time, constraint-based algorithm. It uses tests of variable dependence, independence, and conditional independence to restrict the possible causal relationships between variables. Underlying this technique is the Markov condition [16].

Definition 1 (Markov Condition)

Let A be a node in a causal Bayesian network, and let B be any node that is not a descendant of A in the causal network. Then the Markov condition holds if A and B are independent, conditioned on the parents of A.

Assuming the Markov condition, we can make the following causal claims. For instance, suppose we know that A has no cause. Then if B is dependent on A, B must be caused by A, although possibly indirectly. If we have a third variable C dependent on both A and B, then the three variables lie along a causal chain. Variable A, since it has no cause, is at the head of the chain, but we do not know whether B causes C or vice versa. If, however, A and C become independent conditioned on B, then we can conclude, by the Markov condition, that B causes C.

Now suppose two variables B and C are independent, but each is correlated with A. then B and C are not a causal path, but A is on a causal path with both of them, implying either both are ancestors of A or both are descendants of A. If B and C become dependent when conditioned on A, then by the Markov condition, they cannot be descendants of A, so we can conclude that B and C are causes of A.

In our diabetic application, we modified the LCD algorithm (see Figure 4) to determine causal relationships involving multiple factors. The algorithm assumes tests of dependence and conditional independence using χ^2 statistical test.

```
Input: a set M of rules of the form (A->B), where A contains 2 or more attributes
       conditions. A set N of rules of the form (C->B) where C is a single attribute
       condition.
Output: a list of causal relationships supported by the data
For each rule (A->B) in M
    For each rule (C->B) in N
        If  $C \not\subset A$
            Perform CI(A, B, C): $\chi^2$ (AB|C=0) + $\chi^2$ (AB|C=1)
    EndFor
    If all chi-squares $\geq$threshold $\chi_\alpha^2$
        Output ‘A cause B’
EndFor
```

Figure 4. The Modified LCD Algorithm.

The definitions of dependence and conditional independence are given as follows.

- Dependence test

  Let $s \in (0,1)$ be a support threshold and $c \in (0,1)$ be a confidence threshold. An itemset $S \subseteq I$ is said to be (s,c) correlated if the following two conditions are met:

  1. the value of support(S) exceeds s;
  2. the χ^2 value for the set of items S exceeds the χ^2 value at confidence level c.

- Independence test

  Let $s \in (0,1)$ be a support threshold and $c \in (0,1)$ be a confidence threshold. An itemset $S \subseteq I$ is said to be (s,c)-uncorrelated if the following two conditions are met:

  1. the value of support(S) exceeds s;
  2. the χ^2 value for the set of items S does not exceed the χ^2 value at significance level c.

- Conditional independence test

Variables A and B are independent conditioned on C if p (AB|C) = p (A|C) p (B|C). The chi-squared test for conditional independence looks at the statistic χ^2(AB|C=0) + χ^2(AB|C=1), where χ^2(AB|C=i) is the chi-squared value for the pair A, B limited to data where C = i. If χ^2(AB|C=0) + χ^2(AB|C=1) is greater than the threshold value χ_α^2, then A and B are said to be dependent given condition C at significance level α.

Applying these definitions in our modified LCD algorithm, we obtain the results as shown in Table 2. Our doctors agree that the factors we discovered make sense and that they are helpful to allow the doctors to have better understanding of how multiple attributes interact with each other in the diabetic database.

Table 2. Results of multiple factors analysis.

Year	Number of multiple factors	Sample useful factors

92	3	RACE=1,CLIN=1->Disease=Y
		RACE=1,SEX=1->Disease=Y
		SEX=1,HCAT=Y->Disease=Y

93	13	AGE1>53.5,HCAT=Y-> Disease=Y AGE2>47.5,RACE=1-> Disease=Y ⋮
94	7	AGE1>54.5,RACE=1,SEX=2-> Disease=Y RACE=1,SEX=1->Disease=Y ⋮
96	2	HDUR<9.5,CLIN=6->Disease=Y RACE=1,SEX=1->Disease=Y

However, all the above techniques give only a bird eye's view of the whole domain. In order for the doctors to gain an overview or summary of the disease patterns, we have embarked on the general and exception rules mining (see more details of this technique in [9]). A new representation, called general and exception rule representation, is proposed. This new representation consists of two parts: the general rules and the exception rules. The general rules give the underlying trend in the data while the exception rules give us the abnormalities to these trends. The advantage of this new representation is that it is highly intuitive and can easily draw the user's attention to the interesting rules/patterns. The exception rule mining program takes as input the prunned tree generated by C4.5 [11] and performs the following check to find the exception rules:

- It conducts chi-square test for each node of the tree, if it is greater than the threshold χ_α^2, then it is considered as a general rule candidate.
- If this general rule candidate does not has any parent node that is a general rule node, then it is a top level general rule node; otherwise we need to find the candidate's nearest general-rule parent node, compare its class with the candidate's class label, if they are the same, this is not a sub-general rule node, otherwise it is considered to be a sub-general rule node.
- for each general rule or sub-general rule node, check the leaf node below, if it has different class from the general rule node, then it is considered as an exception node.
- For those nodes that does not belong to the above three categories, we can discard them.

After showing these rules to the doctors, they find the rules useful. In particular, they find the exception rules especially helpful in understanding some of the special subpopulation that exhibits trends which are contrary to the main population. Table 3 gives a summary of the number of general and exception rules discovered. In the year 1993, the decision tree did not produce any significant rules, and hence there is no general rule and exception. From the table, we see that the doctors do have some feelings about the general trends. In fact, the general rules serve to confirm what they know. On the other hand, there are quite a number of exception rules that are unknown to the doctors. They agree that these exception rules are worth investigating further.

Table 3. Results of general and exception rule mining.

Year	Number of General Rules Generated	Number of General Rules Known to the Doctors	Number of Exception rules generated	Number of Exceptions Known to Doctors
92	1	1	2	1
93	0	0	0	0
94	2	2	4	2
95	2	2	8	4
96	2	2	13	3

5. Conclusions

The issues of preprocessing and postprocessing (before and after rule generation) have largely been ignored by the data mining research community. Yet these issues are critical to the success of any real-life applications. To deal with these issues, we have proposed the use of a semi-automatic data cleaning system for cleaning the noisy data and a exploration mining strategy for easy understanding of the rules generated by state-of-the-art data mining techniques. Our doctors confirm that many of the rules discovered conform to the trends that they have observed in their practices. However, they are surprised by some of the exception rules and express interest in investigating them further.

References

[1] Agrawal R., Imielinski T. and Swami, A., “Mining association rules between sets of items in large databases,” ACM *SIGMOD-1993*, 1993, pages 207-216.

[2] Hernandez M. and Stolfo S., “The merge/purge problem for large databases,” *ACM SIGMOD-1995*, 1995, pages 127-138.

[3] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I. “Finding interesting rules from large sets of discovered association rules,” *CIKM-1994.*

[4] Lee M. L., Lu H., Ling T.W. and Ko Y.K., “Cleansing Data for Mining and Warehousing,” *Proceedings of the 10th International Conference on Database and Expert Systems Applications (DEXA99)*, Florence, Italy, August 1999.

[5] Liu B., Hsu W., “Post-analysis of learned rules,” *AAAI-96*, 1996, pp. 828-834.

[6] Liu B., Hsu W., and Chen S., “Using general impressions to analyze discovered classification rules,” *KDD-97*, 1997.

[7] Liu B., Hsu W., and Ma Y., “Integrating classification and association rule mining,” *KDD-98*, 1998.

[8] Liu B., Hsu W., Ma Y., “Pruning and Summarizing the Discovered Associations,” *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-99*, San Diego, USA, 1999, pp. 125-134.

[9] Liu B., Hu, M., and Hsu W., “Representing discovered knowledge as general rules and exceptions.” *Technical report*, School of Computing, National University of Singapore, 2000.

[10] Pazzani M. J., Manu S., Shankle W. R., “Beyond Concise and Colorful: Learning Intelligible Rules,” *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, KDD97,* Newport Beach, USA, 1997, pp 235-238.

[11] Quinlan R., C4.5: A program for machine learning. Morgan Kaufmann, 1992.

[12] Srikant, R., Vu, Q. and Agrawal, R., “Mining association rules with item constraints,” *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, KDD-97*, Newport Beach, USA, 1997, pp 67-73.

[13] Silberschatz A. and Tuzhilin A., “What makes patterns interesting in knowledge discovery systems,” *IEEE Trans. on Know. and Data Eng.*, 8(6), 1996, 970-974.

[14] Tsumoto S., “Automated Discovery of Plausible Rules Based on Rough Sets and Rough Inclusion,” *Proceedings of the Third Pacific-Asia Conference, PAKDD-99*, Beijing, China, 1999, pp 210-219.

[15] Silverstein C., Brin S., Motwani R., Ullman J., “Scalable Techniques for Mining Causal Structures,” Department of Computer Science, Stanford University, 1998.

[16] Spirtes P., Glymour C., and Scheines R. *Causation, Predication, and Search*. Springer-Verlag, New York, 1993.

Acknowledgement

We would like to thank Dr Shanta C Emmanuel, Dr Paul Goh, and Dr Jonathan Phang of the Family Health Service, Ministry of Health, Singapore, for providing us the data as well as their active involvement in the project.