Efficient Discovery of Risk Patterns in Medical Data

Jiuyong Li
School of Computer and Information Science,
University of South Australia,

Ada Wai-chee Fu
Department of Computer Science and Engineering
Chinese University of Hong Kong,

Paul Fahey
Department of Mathematics and Computing
University of Southern Queensland, Australia

Summary:
Objective: This paper studies a problem of efficiently discovering risk patterns in medical data. Risk patterns are defined by a statistical metric, relative risk, which has been widely used in epidemiological research.
Method: To avoid fruitless search in the complete exploration of risk patterns, we define optimal risk pattern set to exclude superfluous patterns, i.e. complicated patterns with lower relative risk than their corresponding simpler form patterns. We prove that mining optimal risk pattern sets conforms an anti-monotone property that supports an efficient mining algorithm. We propose an efficient algorithm for mining optimal risk pattern sets based on this property. We also propose a hierarchical structure to present discovered patterns for the easy perusal by domain experts.
Result: The proposed approach is compared with two well known rule discovery methods, decision tree and association rule mining approaches on benchmark data sets and applied to a real world application. The proposed method discovers more and better quality risk patterns than a decision tree approach. The decision tree method is not designed for such applications and is inadequate for pattern exploring. The proposed method does not discover a large number of uninteresting superfluous patterns as an association mining approach does. The proposed method is more efficient than an association rule mining method. A real world case study shows that the method reveals some interesting risk patterns to medical practitioners.
Conclusion: The proposed method is an efficient approach to explore risk patterns. It quickly identifies cohorts of patients that are vulnerable to a risk outcome from a large data set. The proposed method is useful for exploratory study on large medical data to generate and refine hypotheses. The method is also useful for designing medical surveillance systems.

Keywords: Relative risk, risk pattern, data mining, association rule, decision tree, epidemiology.
Software: http://www.cis.unisa.edu.au/ lijy/Risk-Preventive/Risk-Preventive.html

1 Introduction

1.1 Background and aims

Hospitals and clinics accumulate a huge amount of patient data over the years. These data provide a basis for the analysis of risk factors for many diseases. For example, we can compare cancer patients with non-cancer patients to find patterns associated with cancer. This method has been common practice in *evidence-based medicine*, which is an approach to the practice of medicine in which a clinician is aware of the evidence in support of clinical practice, and the strength of that evidence.

The analysis of the data from comparative studies has usually been done by using statistical software tools, such as SPSS. This is a labor intensive process. It is inefficient to run an exhaustive analysis of interactions of 3 or more exposure variables. Therefore, an automatic data mining tool is required to perform such tedious and time consuming tasks.

The interpretability of results is a requirement for designing a data mining method for medical applications. In general, medical practitioners and researchers do not care how sophisticated a data mining method is, but they do care how understandable its results are.

Rules are a type of the most human-understandable knowledge, and therefore they are suitable for medical applications. There following are two widely used approaches to extract rules from data.

Decision trees, typified by C4.5 [1], can be extended to rules. Decision trees are usually used for building diagnosis models for medical applications [2, 3, 4]. The main objective is to minimise the overall errors in classification. Rules from a decision tree are usually accurate but some are not statistically significant. Furthermore, a decision tree only represents one model among a number of possible models. Rules from a decision tree may fail to present relationships that are of interest to users.

Association rule mining [5] is a general purpose rule discovery scheme. It has been widely used for discovering rules in medical applications [6, 7, 8]. Three challenges of association rule mining approaches in these applications are 1) most widely used interestingness criteria, such as confidence and lift, do not make sense to medical practitioners, 2) too many trivial rules discovered overwhelm truly interesting rules, and 3) an association rule mining approach is inefficient when the frequency requirement, the minimum support, is set low.

To tackle the above problems, we use a widely used epidemiological term, relative risk, to define risk patterns. We propose optimal risk pattern sets to exclude superfluous patterns that are of no interest to medical practitioners. We present an efficient algorithm to discover optimal risk pattern sets. We also study a way to present structured risk patterns to medical practitioners. The proposed method has been applied to a real world application and produced some interesting results. This paper extends our previous work [9].

1.2 Related work

Decision trees are a popular logical method for classification. A decision tree is a hierarchical structure that partitions data into some disjoint groups based on their different attribute values. Leafs of a decision tree contain records of one or nearly one class, and so it has been used for classification. An advantage of decision tree methods is that decision trees can be converted into understandable rules. A most widely used decision tree system is C4.5 [1], its ancestor ID3 [10], and a commercial version C5.0.

Decision trees have been mainly used to build diagnosis models for medical data [2, 3, 4]. When it is used for exploring patterns in medical data, work in [11] shows that it is inadequate for such exploration. One reason is that the objective of decision trees is not to explore data but to build a simple classification model on the data. Another reason is that the heuristic search of decision tree prevents its finding many quality rules. Decision trees only follow one path in tree construction, and hence may miss better rules along alternative paths. Recently, a variant decision tree algorithm, high-yield-partition tree method, has been proposed to discover hi-utility patterns for business intelligence [12]. Its application to medical data is to be explored.

Association rule mining is a major data mining technique, and is a most commonly used pattern discovery method. It retrieves all frequent patterns in a data set and forms interesting rules among frequent patterns. Most frequently used association rule mining methods are Apriori [13] and FP-growth [14].

Association rule mining has been widely used in medical data analysis. Brossette et al. [6] uncovered association rules in hospital infection control and public surveillance data. Paetz et al. [8] discovered association rules in septic shock patient data. Sequential patterns have been found in chronic hepatitis data by Ohsaki et al. [7], and in adverse drug reaction data by Chen et al. [15]. Ordonez et al. used association rules to predict heart disease [16]. However, the discovery of too many rules is a major problem in all applications. Too many trivial and repetitive rules hide truly interesting rules. Association rule mining is inefficient when the frequency requirement, i.e. the minimum support, is set low. Furthermore, the lack of the right interestingness measurements for medical application is another problem.

Some efficient variants of association rule mining have been presented in the last few years, for example, mining non-redundant association rules [17], mining constraint association rules [18], mining most interesting rules [19], mining top N-association rules [20], and mining k-optimal rules [21] or patterns [22]. The rules are defined by confidence, lift or leverage, and hence their results are not directly understandable to medical practitioners. Apart from the first two methods, they have a data coverage problem. For example, the top k rules may come from the same section of data, and this leaves some records in a data set uncovered. As a result, some records in data are not represented in the results.

To our best knowledge, there is only one paper in data mining literature discussing finding patterns defined by relative risk. Li et al. [23] studied a number of algorithms to discover the most general and the most specific patterns defined by relative risk using the convex property of plateaus of support. The most efficient algorithm in [23] is comparable to that of mining minimal generators. We will show theoretically that our approach is more efficient than mining minimal generators in Section 2.2.

2 Method

2.1 Problem definitions

2.1.1 Risk patterns

Let us assume that there is a collection of patient records. Each record is described by a number of discrete attributes, one of which is the target attribute. The target attribute takes two values: abnormal and non-abnormal. Records for patients with a disease or risk under study are labelled as *abnormal*, otherwise records are labelled as *non-abnormal*. An example of such a data set is listed as Table 1.

Gender	Age	Smoking	Blood pressure	...	Class
M	40 - 50	Y	high	...	abnormal
M	20 - 40	N	normal	...	non-abnormal
F	20 - 40	N	normal	...	non-abnormal
⋮	⋮	⋮	⋮	...	⋮

Table 1: An example of medical data set

In the following we refer to the abnormal class as a and the non-abnormal class as n.

A pattern is defined as a set of attribute-value pairs. For example, $\{$Gender = M, Age in [40,50]$\}$ is a pattern with two attribute-value pairs. The *support* of pattern P is the ratio of the number of records containing P to the number of all records in the data set, denoted by $\mathrm{supp}(P)$. When the data set is large, we have $\mathrm{supp}(P) \approx \mathrm{prob}(P)$.

A pattern is usually called frequent if its support is greater than a given threshold. However, in a medical data set, a pattern in the abnormal group would hardly be frequent when the abnormal cases are themselves rare. Therefore, we define the *local support* of P as the support of P in the abnormal group, represented as

$$\mathrm{lsupp}(P \to a) = \frac{\mathrm{supp}(Pa)}{\mathrm{supp}(a)}$$

where Pa is an abbreviation for $P \wedge a$. Others have called this the recall of the rule $(P \to a)$ [24]. We prefer to call it local support since it observes the anti-monotone property of support: the support of a super pattern is less than or equal to the support of its any sub pattern. In this paper, a pattern is *frequent* if its local support is greater than a given threshold.

A risk pattern in this paper refers to the antecedent of a rule with the consequence of abnormal. For the convenience of our discussions, we introduce another important concept for association rules, *confidence*, in the following.

$$\mathrm{conf}(P \to a) = \frac{\mathrm{supp}(Pa)}{\mathrm{supp}(P)}$$

A pattern separates all records into two groups, a group with the pattern and the other without the pattern, e.g., males between 40 and 50 and the rest. Cohorts separated by a pattern and two classes form a contingency table, see Table 2.

	abnormal (a)	non-abnormal (n)	total
P	$\mathrm{prob}(P, a)$	$\mathrm{prob}(P, n)$	$\mathrm{prob}(P)$
$\neg P$	$\mathrm{prob}(\neg P, a)$	$\mathrm{prob}(\neg P, n)$	$\mathrm{prob}(\neg P)$
total	$\mathrm{prob}(a)$	$\mathrm{prob}(n)$	1

Table 2: A contingency table of a pattern and outcomes

Relative risk is a metric often used in epidemiological studies. It is often used to compare the risk of developing a disease of a group people with a certain characteristic to the other group without the characteristic. The *relative risk (RR)* for the cohort with pattern P being abnormal is defined as follows:

$$\begin{aligned}
\mathrm{RR}(P \to a) &= \mathrm{prob}(a|P) / \mathrm{prob}(a|\neg P) \\
&= \frac{\mathrm{prob}(P, a)}{\mathrm{prob}(P)} / \frac{\mathrm{prob}(\neg P, a)}{\mathrm{prob}(\neg P)} \\
&= \frac{\mathrm{supp}(Pa)}{\mathrm{supp}(P)} / \frac{\mathrm{supp}(\neg Pa)}{\mathrm{supp}(\neg P)} \\
&= \frac{\mathrm{supp}(Pa)\,\mathrm{supp}(\neg P)}{\mathrm{supp}(\neg Pa)\,\mathrm{supp}(P)}
\end{aligned}$$

$\neg P$ means that P does not occur. Pa is an abbreviation of $P \wedge a$. $\mathrm{supp}(\neg P)$ is the fraction of all records that do not contain P, and $\neg Pa$ refers to the records containing a but not P.

For example, if $P =$ “smoking”, a = “lung cancer”, and $\mathrm{RR} = 3.0$, then this means that people who smoke are three times more likely to get lung cancer than those who do not.

A relative risk of less than 1 means the group described by the pattern is less likely to be abnormal. A relative risk of grater than 1 means the group described by the pattern is more likely to be abnormal. Confidence interval of relative risk is determined by the numbers in four cells of the contingency table [25]

We give a formal definition of risk patterns using relative risk in the following.

Definition 1 *Risk patterns are patterns whose local support and relative risk are higher than the user specified minimum local support and relative risk thresholds respectively.*

A primitive goal is to find all risk patterns. However, mining all risk patterns suffers two similar problems as association rule mining: too many discovered patterns and low efficiency for low support. Mining optimal risk pattern sets alleviates the problems.

2.1.2 Optimal risk pattern set

Many risk patterns are of no interest to users. For example, we have two patterns, {SEX = M and HRTFAIL = T and LIVER = T} with relative risk 2.3, and {HRTFAIL = T and LIVER = T} with relative risk 2.4. SEX = M in the first pattern does not increase relative risk and hence we say that the first pattern is superfluous. Thus we introduce the optimal risk pattern set to exclude these superfluous patterns.

Definition 2 *A risk pattern set is optimal if it includes all risk patterns except those whose relative risks are less than or equal to that of one of their sub patterns.*

In the above example, the first pattern will not be in the optimal risk pattern set because it is a super set of the second pattern but has lower relative risk.

Optimal pattern set will exclude many superfluous and uninteresting risk patterns, for example, if pattern "symptom = x" is a risk pattern, many patterns, like "gender = m, symptom = x", "gender = f, symptom = x", "gender = m, age = middle age, symptom = x" with the same or a lower relative risk will be excluded from the optimal risk pattern set. Practically, a pattern with a slight improvement in relative risk over its sub patterns is uninteresting. A minimum improvement requirement can be defined by users. The optimal pattern set makes use of the zero minimum improvement. Mining a pattern set with a nonzero minimum improvement can be extended by post-pruning the optimal pattern set.

In the optimal risk pattern set, the relative risk of a super pattern has to be greater than the relative risk of its every sub pattern. Note that the set of records covered by a super pattern is a subset or at most an equal set of the set of records covered by a sub pattern. Therefore, every record in a data set will be covered by a pattern with the highest relative risk. In other words, the optimal pattern set does not include all patterns, but does include patterns with the highest relative risk for all records.

Another important reason for defining the optimal risk pattern set is that it supports a property for efficient pattern discovery. We will present the property in the following section.

2.2 Anti-monotone property of optimal risk pattern sets

In this section, we will prove that optimal risk pattern set satisfies an anti-monotone property, which supports efficient optimal pattern discovery.

We first introduce notation used in the following lemma and corollary. Px is a proper super pattern of P with one additional attribute-value pair x. We use a to stand for class a, and $\neg a$ to stand for a class that is not a. We can use n instead of $\neg a$ for a two-class problem. We use $\neg a$ because conclusions in this section are true for the multiple class problem too. We have $\mathrm{supp}(\neg a) = 1 - \mathrm{supp}(a)$ and $\mathrm{supp}(P\neg a) = \mathrm{supp}(P) - \mathrm{supp}(Pa)$. Furthermore, we have $\mathrm{supp}(\neg(Px)) = 1 - \mathrm{supp}(Px) = [\mathrm{supp}(\neg Px) + \mathrm{supp}(\neg P\neg x) + \mathrm{supp}(P\neg x) + \mathrm{supp}(Px)] - \mathrm{supp}(Px) = \mathrm{supp}(\neg Px) + \mathrm{supp}(\neg P\neg x) + \mathrm{supp}(P\neg x)$.

Lemma 1 *Anti-monotone property for optimal risk pattern sets*
If ($\mathrm{supp}(Px\neg a) = \mathrm{supp}(P\neg a)$) then pattern Px and all its super patterns do not occur in the optimal risk pattern set.

Proof We first present a proof scheme.

Let PQx be a proper super pattern of PQ. $PQx = Px$ and $PQ = P$ when $Q = \emptyset$. To prove the Lemma, we need to show that $\mathrm{RR}(PQx \to a) \leq \mathrm{RR}(PQ \to a)$.

$$\begin{aligned}
\mathrm{RR}(PQ \to a) &= \frac{\mathrm{supp}(PQa)\,\mathrm{supp}(\neg(PQ))}{\mathrm{supp}(\neg(PQ)a)\,\mathrm{supp}(PQ)} \\
&= \frac{\mathrm{conf}(PQ \to a)}{\mathrm{conf}(\neg(PQ) \to a)} \\
&\geq \frac{\mathrm{conf}(PQx \to a)}{\mathrm{conf}(\neg(PQ) \to a)} \qquad (1) \\
&\geq \frac{\mathrm{conf}(PQx \to a)}{\mathrm{conf}(\neg(PQx) \to a)} \qquad (2) \\
&= \mathrm{RR}(PQx \to a)
\end{aligned}$$

We can deduce that $\mathrm{supp}(PQ\neg a) = \mathrm{supp}(PQx\neg a)$ for any Q from $\mathrm{supp}(P\neg a) = \mathrm{supp}(Px\neg a)$.

Next we prove Step (1). Consider $f(y) = \frac{y}{y+\alpha}$ monotonically increases with y when constant $\alpha > 0$ and $\mathrm{supp}(PQ) \geq \mathrm{supp}(PQx) > 0$.

$$\begin{aligned}
\mathrm{conf}(PQ \to a) &= \frac{\mathrm{supp}(PQa)}{\mathrm{supp}(PQ)} \\
&= \frac{\mathrm{supp}(PQa)}{\mathrm{supp}(PQa) + \mathrm{supp}(PQ\neg a)} \\
&= \frac{\mathrm{supp}(PQa)}{\mathrm{supp}(PQa) + \mathrm{supp}(PQx\neg a)} \\
&\geq \frac{\mathrm{supp}(PQxa)}{\mathrm{supp}(PQxa) + \mathrm{supp}(PQx\neg a)} \\
&= \mathrm{conf}(PQx \to a)
\end{aligned}$$

We then prove Step (2). Note that from $\mathrm{supp}(PQ\neg a) = \mathrm{supp}(PQx\neg a)$, we can deduce that $\mathrm{supp}((PQ)\neg x\neg a) = 0$. Another property we shall make use of is that $f(y) = \frac{y-\alpha}{y}$ monotonically increases with y when constant $\alpha > 0$ and $\mathrm{supp}(\neg(PQx)) \geq \mathrm{supp}(\neg(PQ)) > 0$.

$$\begin{aligned}
&\mathrm{conf}(\neg(PQx) \to a) \\
= \; & \frac{\mathrm{supp}(\neg(PQx)a)}{\mathrm{supp}(\neg(PQx))} \\
= \; & \frac{\mathrm{supp}(\neg(PQx)) - \mathrm{supp}(\neg(PQx)\neg a)}{\mathrm{supp}(\neg(PQx))} \\
= \; & \frac{\mathrm{supp}(\neg(PQx)) - (\mathrm{supp}(\neg(PQ)x\neg a) + \mathrm{supp}(\neg(PQ)\neg x\neg a))}{\mathrm{supp}(\neg(PQx))} \\
& \text{(since } \mathrm{supp}((PQ)\neg x\neg a) = 0.\text{)} \\
= \; & \frac{\mathrm{supp}(\neg(PQx)) - \mathrm{supp}(\neg(PQ)\neg a)}{\mathrm{supp}(\neg(PQx))} \\
\geq \; & \frac{\mathrm{supp}(\neg(PQ)) - \mathrm{supp}(\neg(PQ)\neg a)}{\mathrm{supp}(\neg(PQ))} \\
= \; & \frac{\mathrm{supp}(\neg(PQ)a)}{\mathrm{supp}(\neg(PQ))} \\
= \; & \mathrm{conf}(\neg(PQ) \to a)
\end{aligned}$$

The Lemma has been proved. □

From the above lemma, we can adopt a pruning technique as follows: once we observe that any pattern, e.g., Px, satisfying $\operatorname{supp}(Px\neg a) = \operatorname{supp}(P\neg a)$, we do not need to search for its super patterns, e.g., PQx, since they do not occur in an optimal risk pattern set.

Corollary 1 *Closure property*
if ($\operatorname{supp}(Px) = \operatorname{supp}(P)$*) then pattern* Px *and all its super patterns do not occur in the optimal risk pattern set.*

Proof If $\operatorname{supp}(Px) = \operatorname{supp}(P)$, then $\operatorname{supp}(Px\neg a) = \operatorname{supp}(P\neg a)$. Therefore, all its super patterns do not occur in the optimal risk pattern set according to Lemma 1. □

From the above corollary, we can adopt a pruning technique as follows: once $\operatorname{supp}(Px) = \operatorname{supp}(P)$ is observed, we do not need to search for its super patterns, e.g., PxQ since they will not be in the optimal risk set.

This corollary is closely associated with mining minimal generators [26]. P is a *proper generator* of Px when $\operatorname{supp}(Px) = \operatorname{supp}(P)$. P is called a *minimal generator* if there is no $P' \subset P$ such that $\operatorname{supp}(P') = \operatorname{supp}(P)$. According to Corollary 1, a pattern in an optimal risk pattern set has to be a minimal generator. Corollary 1 is a special case of Lemma 1. Lemma 1 disqualifies many minimal generators from being considered to be in the optimal risk pattern set. As a result, mining optimal risk pattern sets does not search all minimal generators, and therefore is more efficient than mining minimal generators.

2.3 Risk pattern mining and presenting

We now discuss how to discover optimal pattern sets efficiently, and how to present risk patterns in a easy to peruse structure. The algorithm makes use of the anti-monotone property to find optimal risk pattern sets efficiently.

2.3.1 MORE algorithm

A näive method to find an optimal risk pattern set undergoes the following three steps. Firstly, discovering all frequent patterns in the abnormal group. Secondly, forming rules using relative risk to replace confidence. Thirdly, post-pruning a large number of uninteresting rules. This procedure is normally inefficient when the minimum support is low.

Our optimal risk pattern mining algorithm makes use of the anti-monotone property to efficiently prune the search space, and this distinguishes it from an association rule mining algorithm.

The efficiency of an association rule mining algorithm lies in its efficient forward pruning of infrequent itemsets. An itemset is frequent if its support is greater than the minimum support. An itemset is potentially frequent only if all its subsets are frequent, and this property is used to limit the number of itemsets to be searched. This anti-monotone property of frequent itemsets makes forward pruning possible.

Lemma 1 and Corollary 1 are used to forwardly prune risk patterns that do not occur in the optimal risk pattern set. When a pattern satisfies the condition of Lemma 1 or Corollary 1, all its super patterns are pruned. Pseudo-code for mining optimal risk pattern sets is presented in the following.

Algorithm 1 *MORE: Mining Optimal Risk pattErn sets*
Input: data set D, the minimum support σ in abnormal class a, and the minimum relative risk threshold θ.
Output: optimal risk pattern set R

Global data structure: l-pattern sets for $1 \leq l$ (An l-pattern contains l attribute-value pairs.)

```
1)  Set $R = \emptyset$
2)  Count support of 1-patterns in the abnormal class
3)  Generate(1-pattern set)
4)  Select risk patterns and add them to $R$
5)  new pattern set $\leftarrow$ Generate(2-pattern set)
6)  While new pattern set is not empty
7)      Count supports of candidates in new pattern set
8)      Prune(new pattern set)
9)      Add patterns with relative risk greater than $\theta$ to $R$
10)     Prune remaining superfluous patterns in $R$
11)     new pattern set $\leftarrow$ Generate(next level pattern set)
12) Return $R$
```

The above algorithm is self-explanatory. We list two important functions as follows.

Function 1 *Generate($(l+1)$-pattern set)*

```
    // Combining
1)  Let $(l+1)$-pattern set be empty set
2)  For each pair of patterns $S_{l-1}p$ and $S_{l-1}q$ in $l$-pattern set
3)      Insert candidate $S_{l-1}pq$ in $(l+1)$-pattern set
    // Pruning
4)      For all $S_l \subset S_{l-1}pq$
5)          If $S_l$ does not exist in $l$-pattern set
6)          Then remove candidate $S_{l-1}pq$
7)  Return $(l+1)$-pattern set
```

Line (5) is implemented by anti-monotone properties of frequent patterns and optimal risk patterns. A non-existing pattern in a l-pattern set is an infrequent pattern or a pattern satisfying Lemma 1 or Corollary 1. They are pruned in the following function.

Function 2 *Prune($(l+1)$-pattern set)*

```
1)  For each pattern $S$ in $(l+1)$-pattern set
2)      If $\mathrm{supp}(Sa)/\,\mathrm{supp}(a) \leq \sigma$ then remove pattern $S$
3)      Else if there is a sub pattern $S'$ in $l$-pattern set
            such that $\mathrm{supp}(S') = \mathrm{supp}(S)$ or $\mathrm{supp}(S'\neg a) = \mathrm{supp}(S\neg a)$
```

```
4)          Then remove pattern $S$
5)  Return
```

Lines (3) and (4) are implemented according to Lemma 1 and Corollary 1. Not only an infrequent pattern but also a pattern satisfying Lemma 1 or Corollary 1 is removed. Both Lemma 1 and Corollary 1 are very effective and the resultant algorithm is more efficient than an association rule mining algorithm.

In the following, we use an example to show how the algorithm works.

B	C	D	E	A
b_1	c	d	e	a
b	c	d_1	e	a
b	c	d	e_1	a
b	c	d	e	a
b	c_1	d	e_2	a
b	c	d_2	e_3	$\neg a$
b	c_2	d_3	e	$\neg a$
b_2	c_3	d	e	$\neg a$

Table 3: The data set of Example 1

Example 1 Consider the data set D in Table 3, and assume $\sigma = 0.4$ and $\theta = 2.0$.

After line (5) in MORE, the 1-pattern set contains $\{b, c, d, e\}$ and the 2-pattern set comprises $\{bc, bd, be, cd, ce, de\}$. Line (8) in MORE calls function Prune(2-candidate set). Pattern bc is pruned because $\text{supp}(bc\neg a) = \text{supp}(c\neg a)$, and pattern de is pruned because $\text{supp}(de\neg a) = \text{supp}(d\neg a)$. After the pruning, the 2-pattern set becomes $\{bd, be, cd, ce\}$. Line (10) in the MORE calls Function Generate (3-pattern set). Candidate bde is generated in line (3) of Function Generator, and then pruned in line (6) of Function Generator because pattern de does not exist in the 2-pattern set. The same procedure repeats on pattern cde. No 3-pattern is generated and hence the program terminated. The output optimal risk pattern set contains $\{c(RR = 2.4), d(RR = 2.4), bd(RR = 2.5), cd(RR = 2.5), ce(RR = 2.5)\}$. An illustration of the searched patterns and output risk patterns by MORE is shown in Figure 1.

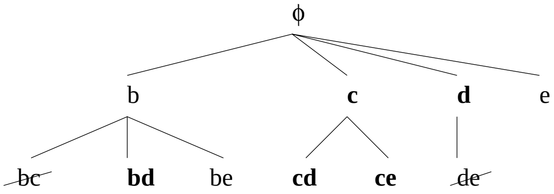


Figure 1: An illustration of the searched patterns and output risk patterns by MORE in Example 1. Patterns crossed are pruned. Patterns in bold are output risk patterns.

As a comparison, we show how to use an association rule mining method to achieve the same goal. We may generate all frequent patterns in class a and form association rules targeting a with the relative risk as the strength. An association rule mining algorithm will examine candidate patterns $\{b, c, d, e, bc, bd, be, cd, ce, de, bcd, bce, cde\}$ and return a set of all risk patterns, $\{c(RR = 2.4), d(RR = 2.4), bd(RR = 2.5), cd(RR = 2.5), ce(RR = 2.5), bcd(RR = 2.0), bce(RR = 2.0), cde(RR = 2.0)\}$. An illustration of the searched patterns and the output risk patterns by an association mining algorithm is shown in Figure 2.

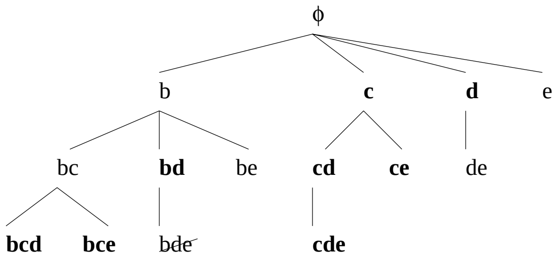


Figure 2: An illustration of the searched patterns and output risk patterns by an association rule mining based approach in Example 1. Only frequent patterns within class a are considered. Patterns crossed are pruned. Patterns in bold are output risk patterns.

We see that the proposed algorithm, MORE, searches a smaller space, and returns a smaller risk pattern set than an association rule mining algorithm. This is a small data set including only a few items. For a large real world data set, differences of the searched spaces and output pattern sets between the two methods are significant.

2.3.2 Pattern presentation

An optimal risk pattern set is smaller than an association rule set, but is still big for medical practitioners to review them. We may only return the top k patterns with the highest relative risk but they may all come from the same section of the data set and lack representatives for all abnormal cases.

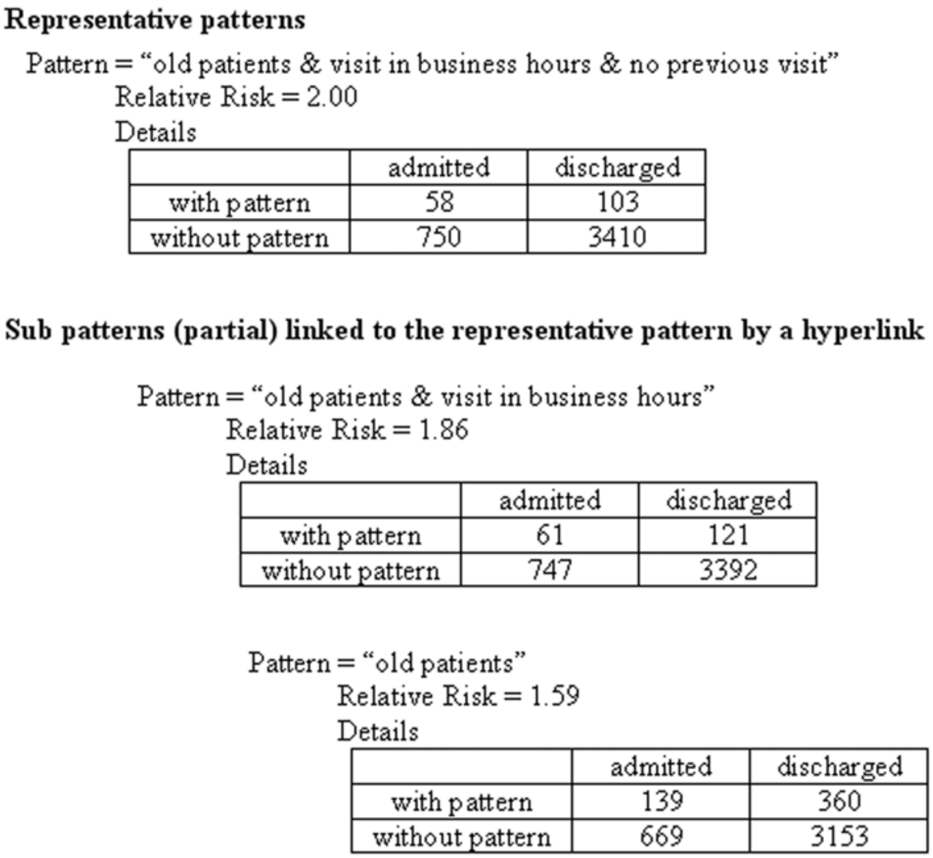
In order to account for all known abnormal cases, we aim to retain a risk pattern with the highest relative risk for each case. We use the following method to select a small set of representative patterns to present to users.

Algorithm 2 *Selecting Representative Risk Patterns*
Input: data set D, and optimal risk pattern set R.
Output: representative risk pattern set R'

```
Set $R' = \emptyset$
For each record $r$ in $D$ belonging to class $a$
    Find all patterns in $R$ that are subsets of $r$
    Add the pattern with the highest relative risk to $R'$
Sort all patterns in $R'$ in the RR decreasing order
Return $R'$
```

As a result, each abnormal record in D has its own representative risk pattern in R'. Through the above selection, the number of patterns becomes manageable.

We organise the remaining risk patterns into a tree structure and hide them behind each representative pattern by using a hyper link. An example is shown in Figure 3. As a result, medical practitioners can easily examine the representative patterns and find their related patterns. This is very useful for finding the evolution of relative risks.



Representative patterns

Pattern = "old patients & visit in business hours & no previous visit"
Relative Risk = 2.00
Details

	admitted	discharged
with pattern	58	103
without pattern	750	3410

Sub patterns (partial) linked to the representative pattern by a hyperlink

Pattern = "old patients & visit in business hours"
Relative Risk = 1.86
Details

	admitted	discharged
with pattern	61	121
without pattern	747	3392

Pattern = "old patients"
Relative Risk = 1.59
Details

	admitted	discharged
with pattern	139	360
without pattern	669	3153

Figure 3: A representative pattern and its sub patterns in a tree structure. Users are presented with a small list of representative patterns. All sub patterns are hidden from users initially, and are brought out when the user clicks the representative pattern in order to know the evolution of the relative risks. There are some repetitions in the tree to make it easy to follow.

3 Experiments and discussions

3.1 Experimental results

Purposes of experiments are to compare MORE to a rule based classification system C5.0, a commercial version of C4.5 [1], and an association rule mining based approach. We used two benchmark medical data sets from UCML repository [27], which are described in Table 4.

Name	#Records	#Attributes	Distributions
Hypothyroid	3163	25	4.8% & 95.2%
Sick	2800	29	6.1% & 93.9%

Table 4: A brief description of data sets used in experiments

3.1.1 Comparison with C5.0

For C5.0, we first used the default setting and then set differential misclassification costs as 20 and 15 for data sets Hypothyroid and Sick respectively. The purpose of differential misclassification costs is to penalise misclassifications in some groups. If we do not set differential misclassification costs (DMC) in Hypothyroid data set, it only results in 4.8% overall error rate that all cases in the hypothyroid group are classified as negative. This small overall error rate reduces the chance of forming rules in hypothyroid group. When we set the differential misclassification costs as 20 in Hypothyroid data set, 1 error in the abnormal group is equivalent to 20 errors in the non-abnormal group. As a result, both types of cases have an equal chance for forming rules.

We set the minimum local support as 5%, and the minimum relative risk as 1.5 for MORE. To compare with C5.0 fairly, we set the maximum number of attribute-value pairs in a pattern as four since most patterns from C5.0 have four or less attribute-value pairs. Rules discovered by C5.0 with the relative risk less than 1.5 and/or with the local support less than 5% are filtered. We use this setting since the number of representative patterns of MORE is comparable to the number of C5.0 rules. If we set the minimum local support low, the number of risk patterns of MORE will be larger. This setting is identical to that in our real world case study in the following section, where setting has been advised by domain experts.

	C5.0		MORE	
data set	default (number)	with DMC (number)	optimal (number)	representative (number)
Hypothyroid	3	5	462	4
Sick	3	7	304	3

Table 5: Comparison with C5.0 by the number of patterns discovered

Table 5 reports the summary of patterns discovered (rules targeting abnormal) of C5.0 and MORE on both data sets. Table 6 lists the average local support and relative risk of discovered patterns by both methods.

Firstly, C5.0 produces fewer patterns than MORE. The total number patterns discovered by MORE is up to 150 times larger than that of C5.0. Setting differential misclassification costs (DMC) does not result in more rules. The exploratory power C5.0 is limited since it discovers few rules. C5.0 is designed for building classification models rather than discovering patterns. In contrast, MORE algorithm is designed for exploring the data to generate hypotheses for further studying. It is not designed for classification.

	C5.0 default		C5.0 with DMC		MORE Representative	
data set	ave(lsupp)	ave(RR)	ave(lsupp)	ave(RR)	ave(lsupp)	ave(RR)
Hypothyroid	0.31	29.4	0.43	23.1	0.78	33.3
Sick	0.40	27.0	0.29	18.6	0.95	43.1

Table 6: Comparison with C5.0 by the quality of discovered patterns using the average local support and relative risk

Secondly, C5.0 fails to find patterns with the highest relative risks. The objective of classification is different from identifying risk patterns. Further, rules discovered by C5.0 tend to be specific and are not supported in data. In contrast, MORE can find patterns with highest relative risk and highest support. This has been demonstrated in Table 6 that both the average local support and the average relative risk of representative patterns are higher than those from C5.0. A fine-tuned decision tree can uncover some interesting patterns in a data set. However, a decision tree does not guarantee the discovery of the patterns with the highest relative risk nor all patterns with the relative risk above a threshold because of its heuristic search trait. The way of search dictates the difference of two methods.

3.1.2 Comparison with variant association rule mining based approaches

Another approach to discover risk pattern sets is based on association rule mining. Firstly, find all frequent patterns in the abnormal group. Secondly, form association rules targeting the abnormal by replacing the confidence with the relative risk. We will show that this approach generates too many patterns and is inefficient in comparison with MORE.

For both MORE and the association rule mining based approach, we set the minimum local support as 5%, the minimum relative risk as 1.5, and the maximum length of patterns as 4. We implemented the association rule mining based approach by Apriori [13]. However, results reported in this section are independent from the implementation since the number of discovered rules and frequent patterns are identical among association rule mining methods. The summary of discovered patterns are listed in Table 7.

data set	Association (pattern number)	MORE (pattern number)
Hypothyroid	21807	462
Sick	24833	304

Table 7: Comparison with an association rule mining approach

The association rule mining approach produces too many patterns and many provide superfluous information. For example, (T3 ≤ 1.15) is a risk pattern because T3 is an indicator for sick. The association rule mining based approach discovers 37 patterns with additional conditions, like (T3 ≤ 1.15, TBGmeasured = f) and (T3 ≤ 1.15, TBGmeasured = f, pregnant= f), which have exactly the same relative risk as pattern

(T3 ≤ 1.15). There are another 4742 patterns containing (T3 ≤ 1.15) which have lower relative risk in the association rule set. All these patterns are not included in the optimal risk pattern set. An optimal risk patterns set is smaller than its corresponding risk pattern set discovered by association rule mining, but includes highest relative risk patterns for all records.

Non-redundant association rule mining [17] makes use of candidates of minimal generators instead of frequent patterns. It avoids generating a lot of superfluous rules and is more efficient than association rule mining. However, non-redundant association rule mining searches all minimal generators that are a superset of candidates searched by MORE. Therefore, non-redundant association rule mining is also less efficient than MORE for mining risk patterns.

To demonstrate the efficiency improvement obtained by MORE over the association rule mining and non-redundant association rule mining approaches, we conducted more experiments using different support settings and high interactions. We searched for risk patterns containing up to ten attribute-value pairs. To make the comparison independent of implementation and computers, we show the number of searched candidates (frequent patterns and minimal generators) instead of the execution time. As a result, the conclusion is general because theoretically the reduction of the searched candidates is the reduction of computational cost. Figure 4 shows that MORE searches fewer candidates than both frequent patterns and minimal generators, and hence is more efficient than the association rule mining and non-redundant association rule mining based approaches. This is more evident when the support is lower.

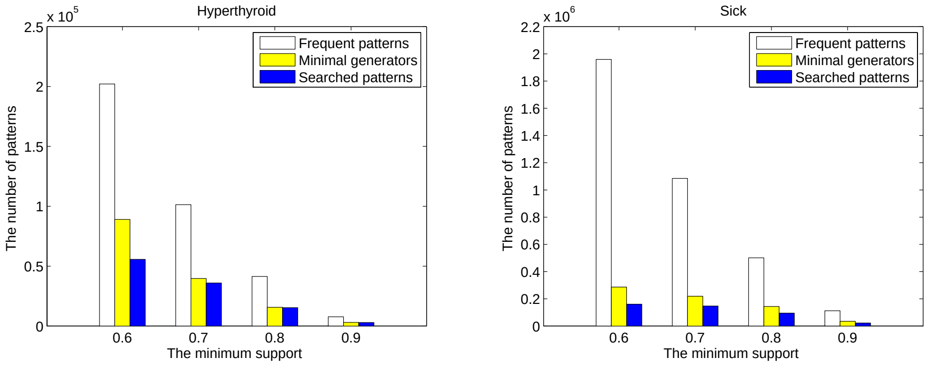


Figure 4: Comparison of the number of patterns searched by MORE with the number of frequent patterns and minimal generators searched by other approaches. MORE searches fewer candidates and hence more efficient.

In sum, C5.0 is not designed for exploring risk patterns. Association rule mining is not efficient for exploring risk patterns, and produces too many risk patterns. MORE is efficient, and produces a manageable number of risk patterns.

3.2 A case study

This method has been applied to a real world applications for analysing emergency department adminstration data.

The hospital in this case study is a regional hospital in Australia. Its emergency department has a 10 bed ultra short stay unit for a short period observation. A patient may stay at the ultra short stay unit for up to 20 hours. Patients staying in the ultra short stay unit may be admitted to the hospital for further treatment, or may be discharged after a brief observation. In some occasions, the beds are not enough to cope with a large demand, and doctors need to transfer some patients to the ward. A significant reduction in administrative work can be achieved if patients who eventually end up at the hospital are admitted to the ward without staying in ultra short stay units after the initial assessment.

The emergency department of Toowoomba Base hospital has collected 4321 patient records who have presented in the ultra short stay unit over two years. 808 records are for patients who were eventually admitted to the hospital and 3513 records are for patients who were discharged after a short stay at ultra short stay units. Doctors are interested in knowing patterns of patients who are admitted. We have done a pilot study on this data set.

Patients are described by 16 attributes. A triage attribute classify patients into 5 groups. Some disease related attributes indicating whether patients have the following problems: renal, cardio, diabetes, and asthma. Some attributes describe personal related information, such as gender, age (categorised into four age groups), marital status, and indigenous status. An attribute indicates location information, in town or off town. Some temporal related attributes show season, month, and week date. An attribute shows whether the patient has visited the hospital within a week. All values are binary or categorical.

We have used C4.5 to analyse the data set firstly. C4.5 builds a model with an accuracy of 82% on this data set. We have not conducted cross validation to evaluate the model since it is not our objective to build a predictive model. Instead, we are interested in the rules discovered by C4.5 targeting admitted class. 20 rules are discovered by C4.5. After we filter rules with 5% local support and 1.5 relative risk thresholds, only two rules are left. Two rules are not enough for doctors to understand the data set. Furthermore, the two rules do not include the pattern with the highest relative risk.

Many rules discovered by C4.5 are of no interest to doctors since the rules do not have sufficient support from data. For example, the first two rules from C4.5 have local supports (in number) of 11 and 3 respectively. They are good classification rules since their confidence are 100%. However, they are of no interest to doctors since they only explain few cases although their relative risk are high. If we consider risk patterns at such low minimum support level, there are thousands of them, i.e. 4105 risk patterns when the minimum number of local support is 9.

Furthermore high accurate rules discovered by C4.5 may not be high relative risk patterns. We show this by the following experiment. We ranked 20 rules discovered by C4.5 first by accuracy and then by relative risk. We calculated the Spearman's rank correlation between the two ranks. Results are shown in Figure 5. We can see that two ranks are loosely correlated. Therefore, discovering accurate classification rules is not suitable for the discovery of risk patterns.

When we applied MORE algorithm to the data set with a support threshold of 5% and relative risk threshold of 1.5, we discovered 131 risk patterns toward admitting to the hospital where 75 patterns are representatives.

Discovered patterns reconfirm many known factors by doctors. For example, patients with cardiovascular or renal related disease are more than two times more likely

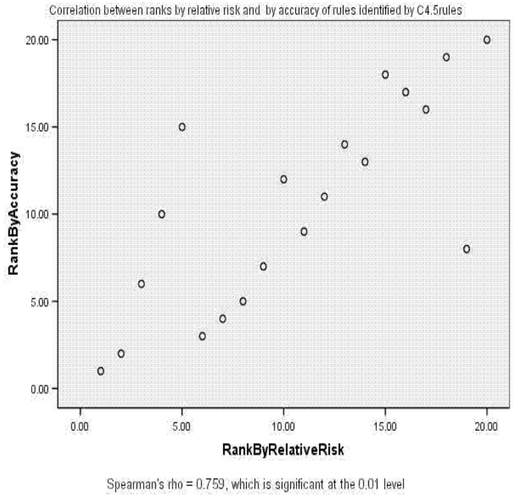


Figure 5: Correlation between the rank by accuracy and the rank by relative risk.Two ranks are loosely correlated. This explains why C4.5 does not find the pattern with the highest relative risk.

to be admitted to the hospital than other patients. Patients with skin/subcutaneous/joint infections are nearly three times more likely to be admitted to the hospital.

Discovered patterns show some common practices used by doctors. Patients who live off town are more likely to be admitted to the hospital even though their situations are not urgent (relative risk of 1.7). This is due to the extra caution of doctors.

Discovered patterns reveal some interesting phenomena. Male patients with limb injuries are nearly two times more likely admitted to the hospital (relative risk of 1.83). Note that neither male patients nor limb injuries alone are risky. This may be attribute to serious injuries in sports or a bias against female limb injury admissions. Patients presented to the department on Mondays in business hours are 1.57 times more likely to be admitted to the hospital than other patients. This shows that the lack of medical service on weekend causes some delayed admissions. Old male patients are very risky (relative risk of 2.23) to be admitted to the hospital. This may be due too the fact that male patients are reluctant to see doctors until there is a pressing urgency.

Since many combinations have been tested in the risk pattern mining process, some patterns becomes significant just by chance. Validation is important to accept or reject them. MORE presents a small set of well structured representative hypotheses quickly from a data set. They can be either validated by domain experts or by further statistical studies. MORE is an efficient data exploratory tool for initial data analysis.

4 Conclusions

This paper has discussed a problem of finding risk patterns in medical data. Risk patterns are defined by an epidemiological metric, relative risk, and hence are understandable to medical practitioners. We define optimal risk pattern set to exclude superfluous patterns that are of no interesting to users. The definition of optimal risk patterns leads to an anti-monotone property for efficient discovery, and we proposed an efficient algorithm for mining optimal risk pattern sets. We have also proposed a way to organise and present discovered patterns to users in an easy to explore structure. The proposed method has been compared with two well known rule discovery methods. The method has also been applied to a real world medical data set and has revealed a number of interesting patterns to medical practitioners. We have the following conclusions from the work: a decision tree approach is unsuitable for discovering risk patterns; an association rule mining approach is inefficient in discovering risk patterns and produces too many uninteresting superfluous patterns; and the proposed algorithm discovers a small set of risk patterns efficiently, which includes the highest relative risk patterns for all records.

The method is useful for exploratory study on large medical data sets. It quickly discovers some "risk spots" in a large medical data set. Results are understandable to medical practitioners. It can be used to generate and refine hypotheses for further time consuming statistical studies.

Acknowledgements

This research has been supported by ARC DP0559090, the RGC Earmarked Research Grant of HKSAR CUHK 4179/01E, and the Innovation and Technology Fund (ITF) in the HKSAR [ITS/069/03]. We thank Dr Mark Laverty for providing the opportunity for the case study.

References

[1] J. R. Quinlan, C4.5: *P*rograms for Machine Learning (Morgan Kaufmann, San Mateo, CA, 1993).

[2] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, *A*rtificial Intelligence in Medicine 1 (2001) 89-109.

[3] J. Li and L. Wong, Using rules to analyse bio-medical data: A comparison between c4.5 and PCL, in: G. Dong, C. Tang, and W. Wang, eds., *A*dvances in Web-Age Information Management, Proceedings of 4th International Conference (Springer, Berlin/Heidelberg, 2003) 254-265.

[4] Z. Zhou and Y. Jiang, Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble, *I*EEE Transactions on Information Technology in Biomedicine 1 (2003) 37-42.

[5] R. Agrawal, T. Imieliński, and A. Swami, Mining association rules between sets of items in large databases, in: P. Buneman and S. Jajodia, eds., *P*roceedings of ACM SIGMOD International Conference on Management of Data (ACM, New York, 1993) 207-216.

[6] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. W. T. Jones, and S. A. Moser, Associarion rules and data mining in hospital infection control and public health

surveillance, Journal of American Medical Informatics Association 5(1998) 373-381.

[7] M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, A rule discovery support system for sequential medical data in the case study of a chronic hepatitis dataset, in: Proceedings of the ECML/PKDD-2003 Discovery Challenge Workshop (http://lisp.vse.cz/challenge/ecmlpkdd2003/, Accessed: 26 June 2008).

[8] J. Paetz and R. W. Brause, A frequent patterns tree approach for rule generation with categorical septic shock patient data, in: J. Crespo, V. Maojo, and F. Martin, eds., Proceedings of the Second International Symposium on Medical Data Analysis (Springer-Verlag, London, 2001) 207-212.

[9] J. Li, A. W. chee Fu, H. He, J. Chen, H. Jin, D. McAullay, G. Williams, R. Sparks, and C. Kelman, Mining risk patterns in medical data, in: R. Grossman, R. J. Bayardo, and K. P. Bennett, eds., Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, New York, 2005) 770-775.

[10] J. R. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81-106.

[11] C. Ordonez, Comparing association rules and decision trees for disease prediction, in: L. Xiong and Y. Xia, eds., Proceedings of the International Workshop on Healthcare Information and Knowledge Management (ACM, New York, 2006) 17-24.

[12] J. Hu and A. Mojsilovic, High-utility pattern mining: A method for discovery of high-utility item sets, Pattern Recognition 11(2007) 3317-3324.

[13] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, in: J. B. Bocca, M. Jarke, and C. Zaniolo, eds., Proceedings of 20th International Conference on Very Large Data Bases (Morgan Kaufmann, San Mateo, CA, 1994) 487-499.

[14] J. Han, J. Pei, Y. Yin, and R. Mao, Mining frequent patterns without candidate generation: A frequent-pattern tree approach, Data Mining and Knowledge Discovery Journal 1(2004) 53-87.

[15] J. Chen, H. He, G. J. Williams, and H. Jin, Temporal sequence associations for rare events, in: H. Dai, R. Srikant, and C. Zhang, eds., Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference (Springer, Berlin/Heidelberg, 2004) 235-239.

[16] C. Ordonez, N. F. Ezquerra, and C. A. Santana, Constraining and summarizing association rules in medical data., Knowledge and Information Systems 3(2006) 1-2.

[17] M. J. Zaki, Mining non-redundant association rules, Data Mining and Knowledge Discovery Journal 3(2004) 223-248.

[18] R. Bayardo, R. Agrawal, and D. Gunopulos, Constraint-based rule mining in large, dense database, *D*ata Mining and Knowledge Discovery Journal 2/3(2000) 217-240.

[19] J. Roberto J. Bayardo and R. Agrawal, Mining the most interesting rules, in: U. Fayyad, S. Chaudhuri and D. Madigan eds., *P*roceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, New York, 1999) 145-154.

[20] G. I. Webb, Efficient search for association rules, in: R. Ramakrishnan, S. Stolfo, R. Bayardo and I. Parsa eds., *P*roceedinmgs of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, New York, 2000) 99-107.

[21] G. I. Webb and S. Zhang, K-optimal rule discovery, *D*ata Mining and Knowledge Discovery Journal 1(2005) 39-79.

[22] Y. Cheung and A. Fu, Mining association rules without support threshold: with and without item constraints, *I*EEE Transactions on Knowledge and Data Engineering 9(2004) 1052-1069.

[23] H. Li, J. Li, L. Wong, M. Feng, and Y.-P. Tan, Relative risk and odds ratio: a data mining perspective, in: C. Li, ed., *P*roceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (ACM, New York, 2005) 368-377.

[24] M. Ohsaki, S. Kitaguchi, K. Okamoto, H. Yokoi, and T. Yamaguchi, Evaluation of rule interestingness measures with a clinical dataset on hepatitis, in: J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, eds., *P*roceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (Springer-Verlag, New York, 2004) 362-373.

[25] M. M. Triola and M. F. Triola, *B*iostatistics for the Biological and Health Sciences (2nd ed.) (Addison-Wesley, Boston, 2005).

[26] J. Wang, J. Han, and J. Pei, Closet+: searching for the best strategies for mining frequent closed itemsets, in: L. Getoor, T. E. Senator, P. Domingos, and C. Faloutsos, eds., *P*roceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, New York, 2003) 236-245.

[27] A. Asuncion and D. J. Newman, UCI repository of machine learning databases (http://archive.ics.uci.edu/ml, Accessed:26 June 2008).