

Mining Risk Patterns in Medical Data

Jiuyong Li
Department of Mathematics
and Computing
University of Southern
Queensland
Toowoomba, Australia, 4350
jiuyong@usq.edu.au

Ada Wai-chee Fu
Department of Computer
Science and Engineering
Chinese University of Hong
Kong
adafu@cse.cuhk.edu.hk

Hongxing He, Jie Chen,
Huidong Jin, Damien McAullay,
Graham Williams¹, Ross Sparks,
Chris Kelman²
CSIRO Mathematical and
Information Sciences
ACT 2601, Australia
firstname.lastname@csiro.au
ATO.graham.williams@togaware.com
NCEPH,ANU,chris.kelman@anu.edu.au

ABSTRACT

In this paper, we discuss a problem of finding risk patterns in medical data. We define risk patterns by a statistical metric, relative risk, which has been widely used in epidemiological research. We characterise the problem of mining risk patterns as an optimal rule discovery problem. We study an anti-monotone property for mining optimal risk pattern sets and present an algorithm to make use of the property in risk pattern discovery. The method has been applied to a real world data set to find patterns associated with an allergic event for ACE inhibitors. The algorithm has generated some useful results for medical researchers.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—*Data mining*; J.3 [Life and Medical Sciences]: Health

General Terms

Algorithm, performance

Keywords

Relative risk, rule, optimal risk pattern set, medical application

1. INTRODUCTION

Over the years hospitals and clinics have collected a huge amount of patient data. These data provide a base 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 practising medicine in which a clinician is aware of the evidence in support of clinical practice, and the strength of that evidence. It is an effective way to generate hypotheses for further study, such as a *randomized controlled*

trial or a *cohort study*. In a randomized controlled study, there are two groups, a treatment group and a control group. The treatment group receives the treatment under investigation, and the control group receives either no treatment or some standard default treatment. Patients are randomly assigned to all groups. A cohort study is a study where patients who presently have a certain condition and/or receive a particular treatment are followed over time and compared with another group without those conditions. The cohort study is used when it is not ethical to assign random patients to a harmful practice, say smoking, for a randomized controlled study. Instead the cohort study will find a group of people who smoke and a group of people who do not, and follow them forward through time to see what health problems they develop. See [1] for more details.

However, the comparison has usually been made by manually operating some data analysis tools, e.g. SPSS. This is a labor intensive process, and the comparison is very difficult to be exhaustive and it is very difficult to apply to high level interactions, for example, combination of 3 or 4 exposure variables. Data mining is a booming and comprehensive research area and a lot of novel methods dealing with large data sets have been proposed in the last decade. There are thousands of publications in data mining, but very few of them focus on applications on medical data. The following are some possible reasons.

Understandability of results Data mining results are typically difficult to interpret, and much effort is necessary for domain experts to turn the results to practical use. In general, users do not care how sophisticated a data mining method is, but they do care how understandable its results are. Therefore, no method is acceptable in practice unless its results are understandable. However, a lot of data mining methods have not achieved this goal yet. For example, it is difficult to interpret results from neural networks.

Decision tree, typified by [14], can be extended to rules [15], and their results are more straightforward to interpret. They have been used to solve classification problems in medical data analysis [10, 19]. However, C4.5 does not work well on the skewed cases in medical data where the normal population greatly outnumbers the population with disease. Other rule based classification methods, e.g. CN2 [7, 6], suffer the same problem.

Amount of results The quantity of output from many data mining method is often unmanageable. For example it is quite impossible for domain experts to review a huge number of association rules. Association rule mining has been used in medical data analysis. Brossette *et al* [4] found association rules in hospital infection

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.
Copyright 2005 ACM 1-59593-135-X/05/0008 ...\$5.00.

Definition 1 Risk patterns are frequent patterns whose relative risks are higher than a threshold.

A risk pattern is in fact a rule targeting the abnormal class. Since we are only concerned with the abnormal class, we omit the target of a rule and call it a pattern.

Our primitive goal is to find all risk patterns in a large data set.

2.2 Classification and association rule mining do not work well

This problem looks like a traditional classification problem, but all existing classification methods, e.g. C4.5 [15], do not work well on these highly skewed data sets. The problem lies with the accuracy measurement for this problem. For example, assume a data set contains 100 abnormal cases and 9900 normal cases. Any noises in the normal class, say 1%, overwhelm all patterns in the abnormal class. Therefore, no accurate rules can be found for the abnormal class. Furthermore, most classification systems employ a default prediction. In this case, setting the default to be normal will give 99% accuracy but this accuracy has no meanings for medical practitioners. Although C4.5 has suggested some remedies for skewed data, from our experiences it is still short of solving the problem.

Another important factor, which is different from classification rule mining, is that doctors or patients are interested in knowing the increase in risk of a certain pattern over cases without the pattern. For example, how much would smoking increase the chance of lung cancer. This is a comparison between the chance of lung cancer in the smoking population versus the chance of lung cancer in the non-smoking population. Conventional classification results would not directly give such an indication.

The primitive goal looks like that of association rule mining, but an association rule mining algorithm is not suitable for this problem. Association rule mining finds rules whose support and confidence are above some minimum thresholds. Rules in the abnormal class are easily ignored since they are lowly supported. Also, it is very difficult to find high confidence rules since confidence is an accuracy measurement and suffers the same problem discussed previously. Further, we are interested in rules that generate patterns of high relative risk instead of high confidence rules.

We may alter an association rule mining algorithm for this purpose. We may restrict the results to patterns that are frequent in the abnormal class only, assuming a support threshold is given for the abnormal class. We may also replace the confidence by the relative risk in association rule mining. However, too many rules from an association rule mining algorithm scare away users, and low efficiency with a low support constraint hinders the users' interaction.

2.3 Optimal risk pattern sets

We follow the track of association rule mining, and will solve two problems: too many rules in the result and low efficiency with a low support constraint.

Many patterns from association rule mining are not of interest to users (since we consider one class only, rules are equivalent to patterns.). For example, we have two patterns, {SEX = M and HRT-FAIL = T and LIVER = T} with relative risk 2.3, and {HRT-FAIL = T and LIVER = T} with relative risk 2.4. SEX = M in the first pattern does not increase relative risk and hence 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.

We are aware that some interesting patterns may not be in the optimal risk pattern set. We use an example to show our points. Suppose that we have the following three patterns:

- (1) PVD = T with RR 3.0,
- (2) SEX = F and PVD = T with RR 2.0 and
- (3) SEX = M and PVD = T with RR 4.0.

Patterns (2) and (3) are very interesting since any record with PVD = T will be explained by one of them. However, pattern (2) is excluded by the optimal risk pattern set.

This is a typical example showing that we need patterns in the whole range of relative risk, both small and large. However, consider that we have generated thousands of patterns. Which patterns should we choose to present to users? Normally, we have to rely on a metric. In our case, it is the high relative risk. As a result, patterns with lower relative risk will be ignored anyway.

One goal of this research is to identify some possible high risk patterns for further studies. For the easy examination by domain experts, the found risk patterns are further reduced to representative patterns by a high relative risk criterion. Therefore, patterns with lower relative risks have no chance to be presented to users.

After a small set of interesting patterns with high relative risk are identified, their relevant patterns with lower relative risk are easily retrieved. For example, assume that pattern (3) is found and identified as an interesting pattern by domain experts. Patterns (2) will be retrieved easily. This has been done in our rule exploration stage.

Therefore, we may focus on the patterns with higher relative risks in pattern generation stage, and ignore the patterns with lower relative risks since otherwise results will be confused.

Our primary goal turns to to find optimal risk pattern sets since it accounts for the major computational cost.

3. ANTI-MONOTONE PROPERTY OF OPTIMAL RISK PATTERN SETS

In this section, we will explore an anti-monotone property to support efficiently mining optimal risk pattern sets.

The following are some notations that are used in the following lemma and corollary.

Px is a proper super pattern of P with one additional attribute-value pair x . To make the result general and be applicable to multiple classes, we use $\neg a$ to stand for classes that are not abnormal. In the two class case shown in the previous section, $\neg a = n$. We have the following relationships: $\text{supp}(\neg a) = 1 - \text{supp}(a)$, $\text{supp}(P\neg a) = \text{supp}(P) - \text{supp}(Pa)$, and $\text{supp}(Px\neg a) = \text{supp}(Px) - \text{supp}(Pxa)$.

Lemma 1 Anti-monotone property

if $(\text{supp}(Px\neg a) = \text{supp}(P\neg a))$ then pattern Px and all its super patterns do not occur in the optimal risk pattern set.

PROOF. We omit proof here because of space limit. \square

From the above lemma, we can adopt a pruning technique as follows: once we observe that any pattern, e.g. Px , satisfying $\text{supp}(Px\neg a) = \text{supp}(P\neg a)$, we do not need to search for its super patterns, e.g. PQx since their relative risks cannot be greater than those of their sub patterns, e.g. PQ . Pattern Px is also removed since $RR(Px) \leq RR(P)$.

The lemma is followed by a corollary.

Industry/Government Track Poster

early. We apply our proposed technique on this problem. In particular we focus on the study to determine how ACE inhibitor usage is associated with Angioedema.

ACE inhibitors are used to treat congestive heart failure (CHF) and high blood pressure (hypertension). ACE inhibitors may also be prescribed to patients after a heart attack or to patients with certain kind of kidney problems, especially with diabetes. Angioedema is a swelling (large welts or weals), where the swelling is beneath the skin rather than on the surface. It is associated with the release of histamine and other chemicals into the bloodstream, and is part of the allergic response. The swelling may occur in the face, neck, and in severe cases may compromise breathing.

Our goal is to identify what types of patients are at risk of Angioedema after taking ACE inhibitors. The data for this task consists of all patients exposed to ACE inhibitors. Class a includes all patients who got Angioedema after taking ACE inhibitors, and class n (or $\neg a$) includes all the other patients taking ACE inhibitors but without Angioedema.

Patients are described by 12 general attributes, such as age, gender, indigenous status, the total number of bed days and the eight hospital diagnosis flags, and 15 pharmaceutical attributes, i.e. 14 ATC (Anatomical and Therapeutic Classification) level-1 drugs and the total number of scripts. Numerical attributes are discretised following the instructions of domain experts.

The data set contains 132000 cases, where only 114 are allergic cases. Therefore, the data set is highly skewed.

We set the minimum local support as 0.05, the maximum number of attribute-value pairs in a risk pattern as 4, and the minimum relative risk as 2.0. The program finished within 1 minute. It returned 417 risk patterns, and 37 representative patterns.

The following are the first three representative patterns with the highest relative risk.

Pattern 1: RR = 3.99

- Gender = Female
- Hospital Circulatory Flag = Yes
- Usage of Drugs in category “Various” = Yes

Pattern 2: RR = 3.82

- Age > 60
- Usage of drugs in category of “Genito urinary system and sex hormones” = Yes
- Usage of drugs in category of “Systematic hormonal preparations” = Yes

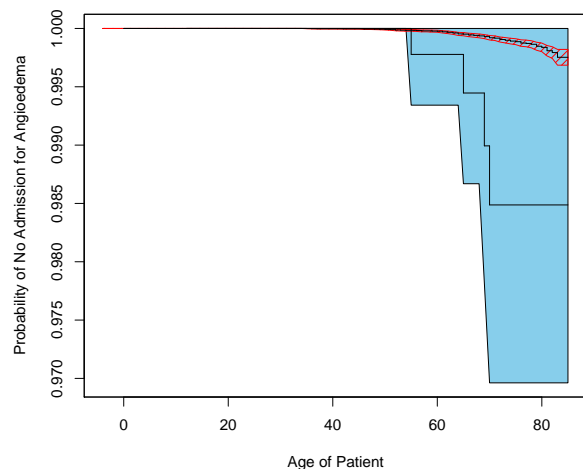
Pattern 3: RR = 3.41

- Usage of drugs in category of “Genito urinary system and sex hormones” = Yes
- Usage of drugs in category of “General anti-infective for systematic use” = Yes
- Usage of drugs in category of “Nervous system” = No

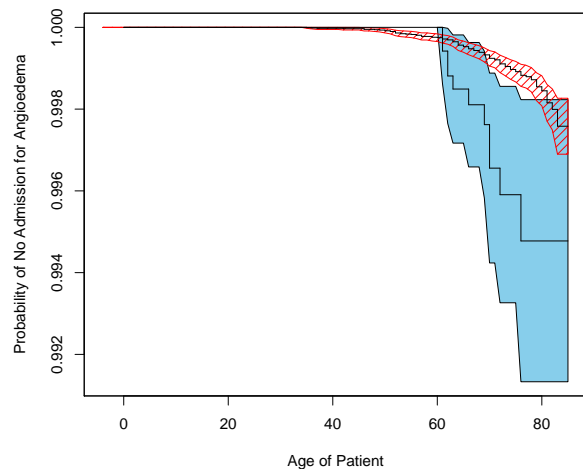
Most found patterns are of great interest to domain experts and verified by them. We have conducted further statistical analysis, e.g. the survival analysis and significance test [16], to evaluate the statistical significance of found patterns.

The survival analysis is concerned with the modelling of ‘lifetime’ data. We estimate the survivor function $S(t)$, by the probability of non-admission to hospitals for Angioedema at age t , to distinguish the subgroup described by the pattern from the others. In addition, we use log-rank test, a formal measure of the strength of evidence that two populations have different lifetimes. It is to detect a difference between groups when the survival curve is consistently higher for one group than another.

Survival Analysis of Pattern 1 (blue one, P-value=4.2229e-09)



Survival Analysis of Pattern 2 (blue one, P-value=2.9257e-05)



Survival Analysis of Pattern 3 (blue one, P-value=2.5506e-09)

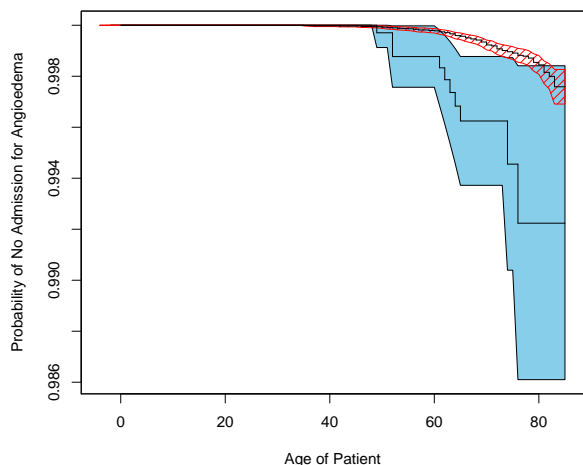


Figure 1: Survival analysis charts of the first three patterns. Blue lines (inside fillings) indicate patient groups identified by the patterns and red lines (inside shades) for the other patients. Fillings and shades show confidence intervals. The groups identified by patterns have significantly higher probability of hospital admission for Angioedema than the other patients for age 50 and above.

