Association Rule Discovery with Unbalanced Class Distributions

Lifang Gu[1], Jiuyong Li[2], Hongxing He[1], Graham Williams[1], Simon Hawkins[1], and Chris Kelman[3]

[1] CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra, ACT 2601, Australia
Firstname.Lastname@csiro.au
[2] Department of Mathematics and Computing
The University of Southern Queensland
jiuyong@usq.edu.au
[3] Commonwealth Department of Health and Ageing
Christopher.Kelman@health.gov.au

Abstract.
There are many methods for finding association rules in very large data. However it is well known that most general association rule discovery methods find too many rules, which include a lot of uninteresting rules. Furthermore, the performances of many such algorithms deteriorate when the minimum support is low. They fail to find many interesting rules even when support is low, particularly in the case of significantly unbalanced classes. In this paper we present an algorithm which finds association rules based on a set of new interestingness criteria. The algorithm is applied to a real-world health data set and successfully identifies groups of patients with high risk of adverse reaction to certain drugs. A statistically guided method of selecting appropriate features has also been developed. Initial results have shown that the proposed algorithm can find interesting patterns from data sets with unbalanced class distributions without performance loss.

Keywords: knowledge discovery and data mining, association rules, record linkage, administrative data, adverse drug reaction.

1 Introduction

The aim of association rule mining is to detect interesting associations between items in a database. It was initially proposed in the context of market basket analysis in transaction databases, and has been extended to solve many other problems such as the classification problem. Association rules for the purpose of classification are often referred to as *predictive association rules*. Usually, predictive association rules are based on relational databases and the consequences of rules are a pre-specified column, called the class attribute.

One of the problems with conventional algorithms for mining predictive association rules is that the number of association rules found is too large to handle,

even after smart pruning. A new algorithm, which mines only the optimal class association rule set, has been developed by one of the authors (Li et al. 2002) to solve this problem.

A second problem with general predictive association rule algorithms is that many interesting association rules are missed even if the minimum support is set very low. This is particularly a problem when a dataset has very unbalanced class distributions, as is typically the case in many real world datasets.

This paper addresses the problem of finding interesting predictive association rules in datasets with unbalanced class distributions. We propose two new interestingness measures for the optimal association rule algorithm developed earlier and use them to find all interesting association rules in a health dataset containing classes which are very small compared to the population.

In collaboration with the Commonwealth Department of Health and Ageing and Queensland Health, CSIRO Data Mining has created a unique cleaned and linked administrative health dataset bringing together State hospital morbidity data and Commonwealth Medicare Benefits Scheme (MBS) and Pharmaceutical Benefits Scheme (PBS) data. The Queensland Linked Data Set (QLDS) links de-identified, administrative, unit-level data, allowing de-identified patients to be tracked through episodes of care as evidenced by their MBS, PBS and Hospital records (Williams et al. 2002). The availability of population-based administrative health data set, such as QLDS, offers a unique opportunity to detect common and rare adverse reactions early. This also presents challenges in developing new methods, which detect adverse drug reactions directly from such administrative data since conventional methods for detecting adverse drug reactions work only on data from spontaneous reporting systems, or carefully designed case-control studies (Bate et al. 1998, DuMouchel 1999, Ottervanger et al. 1998).

This paper presents an association algorithm which uses QLDS to identify groups with high risk of adverse drug reaction. Section 2 describes an algorithm for mining interesting class association rule sets. Section 3 presents a method of feature selection based on statistical analysis. In Section 4 we give a brief description of QLDS and the features selected. Results from mining the optimal class association rule set are then presented in Section 5. Section 6 concludes the paper with a summary of contributions and future work.

2 Interesting Association Rule Discovery

Association rule mining finds interesting patterns and associations among patterns. It is a major research topic in data mining since rules are easy to interpret and understand. However, general association rule discovery methods usually generate too many rules including a lot of uninteresting rules. In addition, most algorithms are inefficient when the minimum support is low. In this section, we propose two new interestingness criteria which overcome problems of existing approaches and can be used in finding interesting patterns in data sets with very unbalanced class distributions.

2.1 Interestingness Criterion Selection

We first introduce the notation used in this paper.
A set of attributes is $\{A_1, A_2, \ldots, A_m, C\}$. Each attribute can take a value from a discrete set of values. For example, $A_i \in \{a_{i_1}, a_{i_2}, \ldots, a_{i_p}\}$. C is a special attribute which contains a set of categories, called classes. A record is $T \in A_1 \times A_2 \times \ldots \times A_m \times C$, and a set of records from a data set is denoted by $D = \{T_1, T_2, \ldots, T_n\}$. A pattern is a subset of a record, $P \subset D$. The support of a pattern P is the ratio of the number of records containing P to the number of all records in D, denoted by $\sup(P)$. A rule is in the form of $P \Rightarrow c$. The support of the rule is $\sup(P \cup c)$ [4]. The confidence of the rule is defined as $\text{conf}(P \Rightarrow c) = \sup(Pc)/\sup(P)$.

If a data set has very unbalanced class distributions many existing interestingness criteria are unable to capture interesting information from the data set. In the following we will use a two-class example to demonstrate this. Assume that $\sup(c_1) = 0.99$ and $\sup(c_2) = 0.01$. Note that $\neg c_1 = c_2$ since there are only two classes.

One of the mostly used interestingness criteria is *support*. A minimum support of 0.01 is too small for class c_1, but too large for class c_2. Another such criterion is *confidence*, which can be written for class c_2 as $\text{conf}(P \Rightarrow c_2) = \frac{\sup(Pc_2)}{\sup(Pc_1)+\sup(Pc_2)}$. We can see that any noise in class c_1 will have a significant impact on class c_2, e.g. causing $\sup(Pc_1) \sim \sup(Pc_2)$. As a result, we can hardly find any high confidence rules in the smaller class, c_2. Similarly, *gain* (Fukuda et al. 1996) suffers the same problem as confidence.

Other alternatives can be similarly evaluated. For example, *conviction* (Brin et al. 1997), defined as $\text{conviction}(P \Rightarrow c) = \frac{\sup(P)\sup(\neg c)}{\sup(P\neg c)}$, measures deviations from the independence by considering outside negation. It has been used for finding interesting rules in census data. A rule is interesting when conviction is far greater than 1. The conviction for class c_2 in our example is written as $\text{conviction}(P \Rightarrow c_2) = \frac{\sup(P)\sup(c_1)}{\sup(Pc_1)}$. Since $\sup(c_1) \approx 1$, we have $\sup(Pc_1) \approx \sup(P)$. As a result, $\text{conviction}(P \Rightarrow c_2) \approx 1$. This means we will not find any interesting rules in the small class using the *conviction* metric.

On the other hand, we can prove that *lift* (Webb 2000), *interest* (Brin et al. 1997), *strength* (Dhar & Tuzhilin 1993)) or the *p-s metric* (Piatetsky-Shapiro 1991) all favour rules in small classes. Here we use lift to show this. The metric *lift* is defined by $\text{lift}(P \Rightarrow c) = \frac{\sup(Pc)}{\sup(P)\sup(c)}$. A rule is interesting if its lift is far greater than 1. For large class c_1 we have $\text{lift}(P \Rightarrow c_1) = \frac{\sup(Pc_1)}{\sup(P)\sup(c_1)} \approx 1$ since $\sup(c_1) \approx 1$. As a result, we can hardly find any high lift rules from the large class using the *lift* metric.

A statistical metric is fair for rules from both large and small classes. One such statistical metric is the *odds-ratio*, which is defined as $\text{Odds-Ratio}(P \Rightarrow c) = \frac{\sup(Pc)\sup(\neg P\neg c)}{\sup((\neg P)c)\sup(P\neg c)}$. However, odds-ratio does not capture the interesting rules we are looking for. We show this with the following examples.

[4] For convenience, we abbreviate $P \cup c$ as Pc in the rest of the paper.

Notation

	P	$\neg P$
c_1	$\sup(Pc_1)$	$\sup(\neg Pc_1)$
c_2	$\sup(Pc_2)$	$\sup(\neg Pc_2)$

Example 1

	P	$\neg P$
c_1	0.297	0.693
c_2	0.006	0.004

Example 2

	P	$\neg P$
c_1	0.792	0.198
c_2	0.0095	0.0005

In Example 1, the probability of pattern P occurring in class c_2 is 0.6, which is twice as that in class c_1. In Example 2, the probability of pattern P occurring in class c_2 is 0.95, which is only 1.19 times of that in class c_1. Hence rule $P \Rightarrow c_2$ is more interesting in Example 1 than in Example 2. However, odds-ratio($A \Rightarrow c_2$) is 3.5 in Example 1 and 4.75 in Example 2 and this leads to miss the interesting rule in Example 1.

To have an interestingness criterion that is fair for all classes regardless of their distributions, we propose the following two metrics.

- Local support, defined in Equation 1:

$$\text{lsup}(P \Rightarrow c) = \sup(Pc)/\sup(c) \tag{1}$$

  When we use local support, the minimum support value will vary according to class distributions. For example, a local support value of 0.1 means 0.099 support in class c_1 and 0.001 in class c_2.
- Exclusiveness, defined in Equation 2:

$$\text{excl}(P \Rightarrow c_i) = \frac{\text{lsup}(P \Rightarrow c_i)}{\sum_j^{|C|} \text{lsup}(P \Rightarrow c_j)} \tag{2}$$

  Such a metric is normalised ($[0, 1]$) and fair for all classes. If pattern P occurs only in class c_i, the exclusiveness will reach one, which is the maximum value.

We now discuss the practical meaning of the exclusiveness metric. From the formula, it can be seen that it is a normalised lift, i.e.,

$$excl(P \Rightarrow c_i) = \frac{\text{lift}(P \Rightarrow c_i)}{\sum_j^{|C|} \text{lift}(P \Rightarrow c_j)} \tag{3}$$

The term lift($P \Rightarrow c_i$) is the ratio of the probability of P occurring in class c_i to the probability of P occurring in data set D. Hence the higher the lift, the more strongly P is associated with class c_i. However as discussed above, lift($P \Rightarrow c_i$) ≈ 1 when $\sup(c_i) \approx 1$. As a result, it is difficult to get a uniform minimum lift cutoff for all classes. From Equation 2, it can be seen that $|C| \times \text{excl}(P \Rightarrow c_i)$ is the ratio of the lift of P in Class c_i to the average lift P in all classes. Therefore, it is possible to set a uniform exclusiveness cutoff value for all classes regardless of their distributions. More importantly, the metric exclusiveness reveals extra information that lift fails to identify. For example, if we have three classes c_1, c_2 and c_3, and $\sup(c_1) = 0.98$, $\sup(c_2) = 0.01$ and $\sup(c_3) = 0.01$, it is possible that we have a pattern P such that both lift($P \Rightarrow c_2$) and lift($P \Rightarrow c_3$) are very high. However, exclusiveness will not be high for either class since P is not exclusive to any single class.

The above two metrics (local support and exclusiveness) and lift are used in our application for identifying groups of high risk to adverse drug reactions.

2.2 Efficient Interesting Rule Discovery

There are many association rule discovery algorithms such as the classic Apriori (Agrawal & Srikant 1994). Other algorithms are faster, including Han et al. (2000), Shenoy et al. (1999), Zaki et al. (1997), but gain speed at the expense of memory. In many real world applications an association rule discovery method fails because it runs out of memory, and hence Apriori remains very competitive. In this application, we do not use any association rule discovery algorithm since it is not necessary to generate *all* association rules for interesting rule discovery.

In the following, we first briefly review definitions of a general algorithm for discovering interesting rule sets and informative rule sets (Li et al. 2001, in press), and then argue that the algorithm fits well with our application.

Given two rules $A \Rightarrow c$ and $AB \Rightarrow c$, we call the latter more specific than the former or the former more general than the latter. Based on this concept, we have the following definition.

Definition 1. *A rule set is an informative rule set if it satisfies the following two conditions:* 1*) it contains all rules that satisfy the minimum support requirement; and* 2*) it excludes all more specific rules with a confidence no greater than that of any of its more general rules.*

A more specific rule covers a subset of records that are covered by one of its more general rules. In other words, a more specific rule has more conditions but explains less cases than any of its more general rules. Hence we only need a more specific rule when it is more interesting than all of its more general rules. Formally, $P \Rightarrow c$ is interesting only if for all $P' \subset P$ Interestingness$(P \Rightarrow c) >$ $Interestingness(P' \Rightarrow c)$. Here Interestingness stands for an interestingness metric. In the current application we use local support, lift and exclusiveness as our metrics. Local support is used to vary the minimum support among unbalanced classes to avoid generating too many rules in one class and too few rules in another class. Lift is used to consider rules in a single small class, and exclusiveness is used for comparing interesting rules among classes. We prove in the following that using the informative rule set does not miss any interesting rules instead of the association rule set.

Lemma 1. *All rules excluded by the informative rule set are uninteresting by the lift and exclusiveness metrics.*

Proof. In this proof we use $AB \Rightarrow c$ to stand for a more specific rule of rule $A \Rightarrow c$. $AB \Rightarrow c$ is excluded from the informative rule set because we have $\text{conf}(AB \Rightarrow c) \leq \text{conf}(A \Rightarrow c)$.

We first prove that the lemma holds for lift. We have
$\text{lift}(A \Rightarrow c) = \frac{\sup(Ac)}{\sup(A)\sup(c)} = \frac{\text{conf}(A \Rightarrow c)}{\sup(c)} \geq \frac{\text{conf}(AB \Rightarrow c)}{\sup(c)} = \text{lift}(AB \Rightarrow c)$
As a result, rule $AB \Rightarrow c$ is uninteresting according to the lift criterion.

For the exclusiveness, we first consider a two-class case, c and $\neg c$. We have $\text{conf}(A \Rightarrow \neg c) = 1 - \text{conf}(A \Rightarrow c)$. Since $\text{conf}(AB \Rightarrow c) \leq \text{conf}(A \Rightarrow c)$, we have

$\text{conf}(AB \Rightarrow \neg c) \geq \text{conf}(A \Rightarrow \neg c)$. Further we have $\text{lift}(AB \Rightarrow c) \leq \text{lift}(A \Rightarrow c)$, $\text{lift}(AB \Rightarrow \neg c) \geq \text{lift}(A \Rightarrow \neg c)$ and therefore

$$\text{excl}(A \Rightarrow c) = \frac{\text{lift}_{(A \Rightarrow c)}}{\text{lift}_{(A \Rightarrow c)} + \text{lift}_{(A \Rightarrow \neg c)}} \geq \frac{\text{lift}_{(A \Rightarrow c)}}{\text{lift}_{(A \Rightarrow c)} + \text{lift}_{(AB \Rightarrow \neg c)}}$$
$$\geq \frac{\text{lift}_{(AB \Rightarrow c)}}{\text{lift}_{(AB \Rightarrow c)} + \text{lift}_{(AB \Rightarrow \neg c)}} = \text{excl}(A \Rightarrow c)$$

Hence, $AB \Rightarrow c$ is uninteresting according to the exclusiveness metric.

For more than two classes, we do not provide a direct proof. Instead we have the following rational analysis. Let c_i be any class. $\sum_j^{|C|} \text{conf}(A \Rightarrow c_j) = \sum_j^{|C|} \text{conf}(AB \Rightarrow c_j) = 1$, When $\text{conf}(AB \Rightarrow c_i) \leq \text{conf}(A \Rightarrow c_i)$, we must have at least one class c_j such that $\text{conf}(AB \Rightarrow c_j) \geq \text{conf}(A \Rightarrow c_j)$. Further, we have $\text{lift}(AB \Rightarrow c_i) \leq \text{lift}(A \Rightarrow c_i)$ and $\text{lift}(AB \Rightarrow c_j) \geq \text{lift}(A \Rightarrow c_j)$. As a result, pattern AB is more interesting in class c_j than in class c_i, and rule $AB \Rightarrow c_i$ is uninteresting by exclusiveness.

Thus we can use the informative rule set instead of the association rule set for interesting rule discovery.

The advantages of using the informative rule set are listed as following. Firstly, the informative rule set is much smaller than the association rule set when the minimum support is small. Secondly, the informative rule set can be generated more efficiently than an association rule set. Thirdly, this efficiency improvement does not require additional memory, and actually the informative (IR) rule set generation algorithm (Li et al. 2001, in press) uses less memory than Apriori.

The IR algorithm was initially implemented on transactional data sets where there are no pre-specified classes. Optimal Class Association Rule set generator (OCAR)(Li et al. 2002) is a variant of the IR algorithm on relational data sets for classification. Since a relational data set is far denser than a transactional data set, OCAR is significantly more efficient than Apriori.

In our application we used a modified OCAR, which uses the exclusiveness as the interestingness metric instead of the estimated accuracy as used in the original OCAR.

Rule discovery requires appropriate features to find significant rules. Feature selection is therefore discussed in the next section.

3 Feature Selection Method

We use statistical methods such as bivariate analysis and logistic regression to identify the most discriminating features associated with patient classes.

3.1 Bivariate Analysis

Assume that the dependent variable represents a yes/no binary outcome, e.g., having the disease or not having the disease, an $m \times 2$ frequency table (Table 1) can be used to illustrate the calculation of the χ^2 value for a categorical

independent variable with m levels. Specifically, we have

$$\chi^2 = \sum_{i=1}^{m} [\frac{(n_{i1} - E_{i1})^2}{E_{i1}} + \frac{(n_{i2} - E_{i2})^2}{E_{i2}}] \tag{4}$$

where E_{ij} is the expected count in cell ij, which is equal to $C_j R_i / N$, $i = 1, \cdots, m$ and $j = 1, 2$. The term, N, refers to the total number of counts.

Independent Variable level	**Dependent Variable**		
	Yes	**No**	**Row Total**
1	n_{11}	n_{12}	R_1
2	n_{21}	n_{22}	R_2
3	n_{31}	n_{32}	R_3
⋮	⋮	⋮	⋮
m	n_{m1}	n_{m2}	R_m
Column Total	C_1	C_2	

Table 1. The $m \times 2$ frequency table. Here the row total, R_i, and the column total, C_j are the sums of all cells in a row and column respectively, i.e., $R_i = n_{i1} + n_{i2}$ and $C_j = \sum_{i=1}^{m} n_{ij}$.

The calculated χ^2 value for each independent variable can be compared with a critical (or cut-off) χ^2 value for $m-1$ degrees of freedom at a required p value. The value p denotes the degree of confidence with which to test the association. For example, a p value of 0.01 indicates that the association will be tested at the 99% confidence. If the calculated χ^2 value is larger than the cut-off value, it can be concluded that the independent variable is associated with the dependent variable in the study population. Otherwise, the independent variable is not related to the dependent variable and will not be selected as a feature.

Similarly if the independent variable is continuous, the t value can be used to test for correlation between the dependent and independent variables. Since our class association rule discovery algorithm only takes categorical variables, calculation details of the t value are skipped.

3.2 Logistic Regression

An alternative multivariate statistical method is logistic regression. A logistic regression model can be written as the following equation:

$$\ln(\frac{p}{1-p}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{5}$$

where p is the probability, at which one of the binary outcomes occurs (e.g., the probability of having the disease), α is the intercept, and β_i is the coefficient of the independent variable x_i. The coefficient, β_i, can be transformed into a more meaningful measure, the odds ratio (OR), by the following formula:

$$OR_i = e^{\beta_i} \tag{6}$$

Odds ratios can be used to infer the direction and magnitude of the effects of the independent variables on the probability of the outcome. An odds ratio greater than 1 indicates that the probability of the outcome (e.g., having the disease) will increase when a continuous independent variable increases a unit in its value or when a specific group of a categorical independent variable is compared to its reference group. For example, if the dependent variable is the *probability of having migraine*, the independent variable is *gender*, and *male* is used as the reference group, an odds ratio of 2.0 then implies that females are 2.0 times more likely to have migraine than males. As a result, only independent variables with odd ratio values much larger or smaller than 1 will be selected as important features.

4 Data

The Queensland Linked Data Set (Williams et al. 2002) has been made available under an agreement between Queensland Health and the Commonwealth Department of Health and Ageing (DoHA). The data set links de-identified patient level hospital separation data (1 July 1995 to 30 June 1999), Medicare Benefits Scheme (MBS) data, and Pharmaceutical Benefits Scheme (PBS) data (1 January 1995 to 31 December 1999) in Queensland.

Each record in the hospital data corresponds to one inpatient episode. Each record in MBS corresponds to one MBS service for one patient. Similarly, each record in PBS corresponds to one prescription service for one patient. As a result, each patient may have more than one hospital, or MBS or PBS record.

4.1 Selection of Study Population

PBS data in QLDS contain mostly prescription claims for concessional or repatriate cardholders. There are a total of 733,335 individuals in PBS and 683,358 of them appear as always concessional or repatriate during our five year study period. This represents 93% of all individuals in PBS. Since the drug usage history of these 683,358 individuals is covered more completely in PBS, a subset of them is chosen as our study population, i.e, those who take a particular type of drug.

The adverse drug reaction to be investigated in this study is angioedema associated with the use of ACE inhibitors (Reid et al. 2002). This is a known adverse reaction and the aim of this investigation is to confirm its existence from administrative health data using the proposed algorithm. Drugs are identified in PBS using the WHO codes, which are based on the Anatomical and Therapeutic Classification (ATC) system. Adverse events are identified in hospital data using principal diagnosis codes. Table 2 shows the number of ACE inhibitor users split into two classes, those having and not having angioedema. It can be seen that the distribution of the two classes is very unbalanced and Class 1 is only 0.088% of the whole study population. However, this is the class we are interested in characterising. Section 2 has already described how to find rules to characterise such an under-represented class.

	Angioedema		
	Yes (Class 1)	No (Class 0)	Total
Counts	116	131,184	132,00
Percentage	0.088	99.912	100

Table 2. Study population split into two classes.

4.2 Feature Selection

The aim of the feature selection process is to select those variables which are strongly associated with the dependent variable, based on statistical analysis described in Section 3. For our particular study the dependent variable is the probability of having angioedema for ACE inhibitor users.

From the hospital data we initially extract variables, such as age, gender, indigenous status, postcode, the total number of bed days, and 8 hospital flags. The last two variables are used to measure the health status of each patient. From the PBS data, 15 variables (the total number of scripts of ACE inhibitors and 14 ATC level-1 drug flags) are initially extracted. The variable "total number of scripts" can be used to indicate how long an individual is on ACE inhibitors. The ATC level-1 drug flags are used to investigate adverse reactions caused by possible interactions between ACE inhibitors and other drugs.

Using the feature selection method described in Section 3, the 15 most discriminating features are selected. These features include age, gender, hospital flags, and flags for exposure to other drugs. The optimal class association rule discovery algorithm then run on the extracted data.

5 Results

The interesting association rule discovery algorithm described in Section 2 is applied to the data set with 132,000 records. There are several input parameters to the algorithm. Two of them are the minimum local support and the maximum length (number of variables used in each rule). In our tests we set the minimum local support to 0.05 and the maximum length to 6. The rules (thousands) are sorted by their value of interestingness in descending order. Only the top few rules with highest interestingness values are described here for verbosity. The top three rules and their characteristics are shown in Table 3.

Rule 1
- Gender = Female
- Age $\geq$ 60
- Took genito urinary system and sex hormone drugs = Yes
- Took Antineoplastic and immunimodulating agent drugs = Yes
- Took musculo-skeletal system drugs = Yes

Rule 2
- Gender = Female
- Had circulatory disease = Yes
- Took systemic hormonal preparation drugs = Yes
- Took musculo-skeletal system drugs = Yes
- Took various other drugs = Yes

Rule 3
- Gender = Female
- Had circulatory disease = Yes
- Had respiratory disease = Yes

	Rule 1	Rule 2	Rule 3
Number of patients in the group	1,549	1,629	1,999
Percentage of this group	1.17	1.23	1.51
Number of patients in Class 0	1,542	1,622	1,991
Number of patients in Class 1	7	7	8
Local support in Class 1	6.03%	6.03%	6.90%
Interestingness	0.838	0.831	0.820
Lift of Class 1	5.14	4.89	4.55

Table 3. Characteristics of groups identified by Rules 1 to 3.

– Took systemic hormonal preparation drugs = Yes
– Took various other drugs = Yes

For example, the group identified by Rule 1 has a lift value of 5.14 for Class 1. This indicates that individuals identified by this rule are 5.14 times more likely to have angioedema than the average ACE inhibitor users.

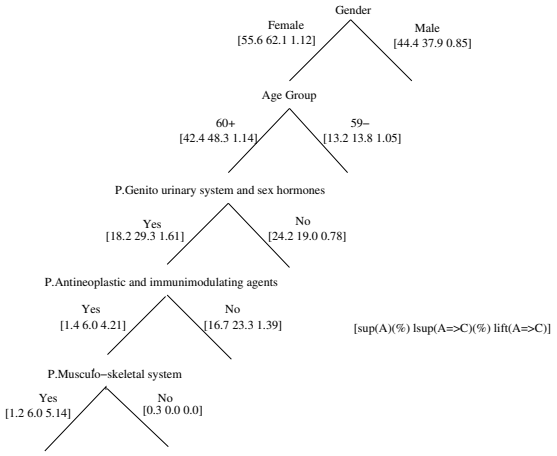

Fig. 1. Probability tree of Rule 1.

Figure 1 provides graphic presentation of Rule 1 in terms of a probability tree. According to the probability tree, female ACE inhibitor users are 1.12 times more likely to have angioedema than the average ACE inhibitor users. For female ACE inhibitor users aged 60 years or older, the likelihood increases to 1.14. As the tree goes further down, i.e., females aged 60 years or older who have also taken genito urinary system and sex hormone drugs, the likelihood increases further to 1.61. If individuals who have all the characteristics mentioned above and have also used antineoplastic and immunimodulating agent drugs, the likelihood goes up to 4.21 times of the average ACE inhibitor users. Finally, in addition to all the above mentioned characteristics, the factor of exposure to musculo-skeletal

system drugs, raises the risk factor of having angioedema up to 5.14 times of the average ACE inhibitor users.

6 Discussion and Conclusions

We have developed an algorithm for mining optimal class association rule sets and have applied the algorithm to a very unbalanced data set to identify groups with high risks of having adverse drug reactions. A feature selection method based on statistical analysis has also been developed to select appropriate features for our data mining algorithm.

Results from testing the association of angioedema with usage of ACE inhibitors have identified groups that are, on average, 4 to 5 times more likely to have an adverse reaction (angioedema) than the average ACE inhibitor users. We note that while the data mining approach has successfully identified key areas in the data worthy of exploration and explanation, conclusions relating to the suitability of ACE inhibitor usage for particular populations need to be further investigated and confirmed by medical specialists, or existing medical studies.

This study has focused on analysing a known adverse reaction to confirm the approach. The more challenging task of identifying unknown adverse reactions from administrative health data like QLDS is ongoing.

Acknowledgements

The authors wish to acknowledge the Commonwealth Department of Health and Ageing and Queensland Health for providing data suitable for linking. Without this data studies of the overall delivery of health care in Australia would not be possible.

References

Agrawal, R. & Srikant, R. (1994), Fast algorithms for mining association rules in large databases, in *Proceedings of the Twentieth International Conference on Very Large Databases*, Santiago, Chile, pp. 487–499.

Bate, A., Lindquist, M., Edwards, I., Olsson, S., R.Orre, Landner, A. & Freitas, R. D. (1998), 'A bayesian neural network method for adverse drug reaction signal generation', *European Journal of Clinical Pharmacology* **54**, 315–321.

Blake, E. K. C. & Merz, C. J. (1998), 'UCI repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html'.

Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. (1997), Dynamic itemset counting and implication rules for market basket data, in *Proceedings, ACM SIGMOD International Conference on Management of Data: SIGMOD 1997: May 13–15, 1997, Tucson, Arizona, USA*, Vol. 26(2), ACM Press, NY, USA, pp. 255–264.

Dhar, V. & Tuzhilin, A. (1993), 'Abstract-driven pattern discovery in databases', *IEEE Transactions on Knowledge and Data Engineering* **5**(6).

DuMouchel, W. (1999), 'Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system', *American Statistical Association* **53**(3), 177–190.

Fukuda, T., Morimoto, Y., Morishita, S. & Tokuyama, T. (1996), Data mining using two-dimensional optimized association rules: scheme, algorithms, and visualization, in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4–6, 1996*, ACM Press, New York, pp. 13–23.

Han, J., Pei, J. & Yin, Y. (2000), Mining frequent patterns without candidate generation, in *Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'00)*, May, pp. 1–12.

Li, J., Shen, H. & Topor, R. (2001), Mining the smallest association rule set for prediction, in *Proceedings of 2001 IEEE International Conference on Data Mining (ICDM 2001)*, IEEE Computer Society Press, pp. 361–368.

Li, J., Shen, H. & Topor, R. (2002), 'Mining the optimal class assciation rule set', *Knowledge-based Systems* **15**(7), 399–405.

Li, J., Shen, H. & Topor, R. (in press), 'Mining informative rule set for prediction', *Journal of intelligent information systems* .

Ottervanger, J., Valkenburg, H. A., Grobbee, D. & Stricker, B. C. (1998), 'Differences in perceived and presented adverse drug reactions in general practice', *Journal of Clinical Epidemiology* **51**(9), 795–799.

Piatetsky-Shapiro, G. (1991), Discovery, analysis and presentation of strong rules, *in* G. Piatetsky-Shapiro, ed., *Knowledge Discovery in Databases*, AAAI Press / The MIT Press, Menlo Park, California, pp. 229–248.

Reid, M., Euerle, B. & Bollinger, M. (2002), 'Angioedema'.
URL: *http://www.emedicine.com/med/topic135.htm*

Shenoy, P., Haritsa, J. R., Sudarshan, S., Bhalotia, G., Bawa, M. & Shah, D. (1999), Turbo-charging vertical mining of large databases, in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD-99)*, ACM SIGMOD Record 29(2), ACM Press, Dallas, Texas, pp. 22–33.

Webb, G. I. (2000), Efficient search for association rules, in *Proceedinmgs of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-00)*, ACM Press, N. Y., pp. 99–107.

Williams, G., Vickers, D., Baxter, R., Hawkins, S., Kelman, C., Solon, R., He, H. & Gu, L. (2002), Queensland linked data set, Technical Report CMIS 02/21, CSIRO Mathematical and Information Sciences, Canberra. Report on the development, structure and content of the Queensland Linked Data Set, in collaboration with the Commonwealth Department of Health and Ageing and Queensland Health.

Zaki, M. J., Parthasarathy, S., Ogihara, M. & Li, W. (1997), New algorithms for fast discovery of association rules, in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, AAAI Press, p. 283.