Temporal Sequence Associations for Rare Events*

Jie Chen, Hongxing He, Graham Williams, and Huidong Jin

CSIRO Data Mining
GPO Box 664, Canberra ACT 2601, Australia
{Jie.Chen,Hongxing.He,Graham.Williams,Warren.Jin}@csiro.au

Abstract. In many real world applications, systematic analysis of rare events, such as credit card frauds and adverse drug reactions, is very important. Their low occurrence rate in large databases often makes it difficult to identify the risk factors from straightforward application of associations and sequential pattern discovery. In this paper we introduce a heuristic to guide the search for interesting patterns associated with rare events from large temporal event sequences. Our approach combines association and sequential pattern discovery with a measure of risk borrowed from epidemiology to assess the interestingness of the discovered patterns. In the experiments, we successfully identify a known drug and several new drug combinations with high risk of adverse reactions. The approach is also applicable to other applications where rare events are of primary interest.

1 Introduction

The present work is motivated by the specific domain of temporal data mining in health care, especially adverse drug reactions. Adverse drug reactions occur infrequently but may lead to serious or life threatening conditions requiring hospitalisation. Thus, systematic monitoring of adverse drug reactions is of financial and social importance. The availability of a population-based prescribing data set, such as the Pharmaceutical Benefits Scheme (PBS) data in Australia, linked to hospital admissions data, provides an opportunity to detect common and rare adverse reactions at a much earlier stage. The problem domain has the following characteristics: (1) Primary interest lies in rare events amongst large datasets; (2) Factors leading to rare adverse drug reactions include temporal drug exposure; (3) Rare events are associated with a small proportion of patients yet all data for all patients are required to assess the risk.

For adverse drug reactions, we usually have little prior knowledge of what drug or drug combinations might lead to unexpected outcomes. Our aim is to

* The authors acknowledge the valuable comments from their colleagues, including C. Carter, R. Baxter, R. Sparks, and C. Kelman, as well as the anonymous reviewers. The authors also acknowledge the Commonwealth Department of Health and Ageing, and the Queensland Department of Health for providing data for this research.

H. Dai, R. Srikant, and C. Zhang (Eds.): PAKDD 2004, LNAI 3056, pp. 235–239, 2004.

discover patterns associated with rare events that are then further assessed for their possible relationship with adverse outcomes. Different from some previous work on mining interesting patterns for group difference [2,3] and health data [8, 4], we propose an approach which extends association and sequential pattern discovery with a heuristic measure motivated from epidemiology. We discover patterns that have high local support but generally low overall support and assess their significance using an estimate of a measure of risk ratio. The paper is organised as follows. Formal definitions and our methods are presented in Section 2. Section 3 reports on some encouraging results. The conclusion and discussion are given in Section 4.

2 Mining Temporal Association for Rare Events

Consider a collection of entities ϵ_i $(i = 1, 2, \ldots) \in E$, for example, credit card holders of a bank or patients in hospital. The activities of each entity ϵ_i are recorded as an event sequence $s_i = \langle (e_{i1}, t_{i1}), (e_{i2}, t_{i2}), \ldots, (e_{ij}, t_{ij}), \ldots, (e_{in_i}, t_{in_i}) \rangle$, where n_i is the number of events for the ith entity. For each event (e_{ij}, t_{ij}), e_{ij} indicates an **event type** and the **timestamp** t_{ij} indicates the time of occurrence of the event. For example, the following sequence describes a set of medical services received by a patient:

$$\langle (\text{G03CA}, 1), (\text{J01DA}, 7), (\text{C08CA}, 10), (\text{C09AA}, 10), (\textit{Angioedema}, 30) \rangle .$$

On day 1 the patient was dispensed the drug *estrogen*, whose ATC code is G03CA. They then took *cephalosporins and related substances* (J01DA) on the 7^{th} day, and *dihydropyridine derivatives* (C08CA) and *ace inhibitor* (C09AA) on the 10^{th} day. Twenty days later they were hospitalised due to *angioedema*. We refer to this particular event of interest as the **target event**, which can be either within the studied event sequence of an entity or just an external event not included in the event sequence but associated with the entity.

Using the target event as a classification criterion, we partition the entities into two subsets. The first one, denoted by $T \subset E$, contains entities having at least one target event. The second one, $\overline{T} \subset E$, consists of all remaining entities. Note that the time spans can be quite long in any particular sequence. Quite often, only events occurring within a particular lead up period, prior to the target event, are relevant. We first define a time window and its associated segments as follows.

Definition 1. *$[t_s, t_e]$ is a **time window** that starts at time t_s and ends at time t_e, where $w = t_e - t_s$ is constant, and usually specified by a domain expert.*

Definition 2. *$\langle (e_{ip}, t_{ip}), (e_{i,p+1}, t_{i,p+1}), ..., (e_{iq}, t_{iq}) \rangle$ is a **windowed segment** of sequence s_i with time window $[t_s, t_e]$ if $t_s \leq t_{ip} \leq t_{i,p+1}, \ldots, t_{iq} < t_e \leq t_{in_i}$, $t_{i,p-1} < t_s$ and $t_{i,q+1} \geq t_e$.*

Definition 3. *For any target entity $\epsilon_i \in T$, a **target segment** of s_i is a windowed segment of s_i with time window $[t_s, t_e]$ where t_e indicates the first occurrence time of the target event.*

Definition 4. *For any entity* $\epsilon_i \in \overline{T}$*, a* ***virtual target segment*** *of* s_i *is a windowed segment of* s_i *with time window* $[t_s, t_e]$*.*

There may be more than one target event associated with an entity in T. For simplicity, only the first target event is considered. Medical advice suggested that only events occurring within some time window prior to the target event might be considered in this exploration. For example, drug usage within six month prior to the adverse reaction is of main interest in our application. Given the fixed window length w, we have a list of virtual target segments for entity ϵ_i with different starting timestamps, $t_{i1}, t_{i2}, ..., \text{or}, t_{il_i}$, where t_{il_i} is the first element in $\langle t_{i1}, t_{i2}, ..., t_{in_i} \rangle$ such that $t_{il_i} \geq t_{in_i} - w$. We denote all these l_i virtual target segments as $L(i, w)$. Also, we can prove that any non-empty virtual target segment with $t_s \in R^1$ must be in $L(i, w)$.

We introduce a risk ratio, as often used in epidemiological studies, to measure the association between a factor and a disease, i.e., being a ratio of the risk of being disease positive for those with and those without the factor [1, p672]. We use the following estimate of a risk ratio for a candidate pattern p occurring within a fixed sized time window:

$$RR(p, w) = \frac{|T| s_T(p)}{|T| s_T(p) + |\overline{T}| s_{\overline{T}}(p)} \Big/ \frac{|T|(1 - s_T(p))}{|T|(1 - s_T(p)) + |\overline{T}|(1 - s_{\overline{T}}(p))}. \quad (1)$$

where $s_T(\text{p})$ is the **support in** T of pattern p, defined as the proportion of entities in T having p in their target segments, and $s_{\overline{T}}(\text{p})$ is the **support in** $\overline{T}$ of p, the proportion of entities in $\overline{T}$ containing p in their virtual target segments, i.e., p is contained in any element of $L(i, w)$. A risk ratio of 1 (i.e., $RR(p, w) = 1$) implies that there is equal risk of the disease with or without pattern p within a fixed sized time window. When $RR(p, w) > 1$, there is a greater risk of the disease in the exposed group.

We employ both the support in T and the estimated risk ratio for identifying interesting patterns. The framework for mining interesting patterns associated with rare events includes the following steps. Firstly, we extract two datasets of entities in the problem domain. Each entity records demographic data and an event sequence. The first dataset contains all entities with at least one target event. The second dataset contains the other entities. Secondly, we partition the entities of the two datasets into sub-populations according to their demographics. Thirdly, we discover candidate association [6] and sequential patterns [5] in T of each sub-population. The events within a fixed sized time window prior to the target event are used for pattern discovery. Fourthly, we explore corresponding $\overline{T}$ of each sub-population to identify patterns mined in the above step. Finally, estimated risk ratios of the candidate patterns for each sub-population are calculated according to Equation 1.

3 Experimental Results

The Queensland Linked Data Set [7] links hospital admissions data from Queensland Health with the pharmaceutical prescription data from Commonwealth Department of Health and Ageing, providing a de-identified dataset for analysis.

Table 1. Estimated risk ratio of sample discovered association patterns. $|T|/|\overline{T}|$ for these cohorts are 55/104257 (Male 20-59), 76/194789 (Female 20-59), and 73/128586 (Female 60+)

Gender	Age	Pattern	support %	chi-square	RR
Male	20-59	C09AA N06AA	9.0	8.896	3.684
Male	20-59	N06AA H02AB	9.0	5.613	2.888
Female	20-59	C09AA G03CA	9.2	8.991	3.090
Female	60+	C09AA G03CA	24.6	19.45	3.112
Female	60+	C09AA C08CA	26.0	4.435	1.741

The record for each patient includes demographic variables and a sequence of PBS events for a five year period. Two datasets are extracted. One contains all 299 patients with hospital admissions due to angioedema, e.g., target event. The other contains 683,059 patients who have no angioedema hospitalisations.

It should be noted that the studied population consists of hospital patients rather than the whole Queensland population. We make the assumption that prior to the period of study these patients were not hospitalised for angioedema.

Tables 1 and 2 list experimental results for particular age/gender cohorts. The fourth column lists the support for the pattern in the target dataset. The other columns show the chi-square value and the estimated risk ratio, respectively. Here we use the chi-squares value, which is calculated together with the estimated risk ratio, as a threshold to constrain resulting patterns. Thus, for males aged 20-59 the drugs C09AA and N06AA within six months are over three times more likely to be associated with angioedema patients than with the non-angioedema patients. The following are some of the interesting sequential patterns. Usage of *estrogen* (G03CA) followed by *ace inhibitor* (C09AA), which is a known drug possibly resulting in angioedema, within six months is associated with the estimated risk ratio 2.636 of angioedema for females aged 60+; The sequence consisting of *dihydropyridine derivatives* (C08CA) and *ace inhibitor* (C09AA) within six months is generally associated with a high estimated risk ratio of angioedema for females aged 60+. Interestingly, four of these sequences begin with the pair of C08CA and C09AA, which means the two drugs are supplied to patients on the same day. These proposed hypotheses then form the basis for further statistical study and validation.

Table 2. Estimated risk ratio of sequential patterns. $|T|/|\overline{T}|$ for these cohorts are 76/194789 (Female 20-59), and 73/128586 (Female 60+).

Gender	Age	Pattern	support %	chi-square	RR
Female	20-59	C09AA -1 C09AA	15.7	7.210	2.273
Female	60+	G03CA -1 C09AA	20.5	12.11	2.636
Female	60+	C08CA C09AA -1 C08CA -1 C09AA -1 C08CA	17.8	9.207	2.453
Female	60+	C08CA C09AA -1 C09AA -1 C08CA -1 C08CA	17.8	9.053	2.436
Female	60+	C08CA -1 C09AA -1 C09AA -1 C08CA	20.5	9.395	2.365
Female	60+	C09AA -1 C08CA -1 C09AA -1 C08CA -1 C09AA	19.1	8.751	2.346
Female	60+	C08CA C09AA -1 C09AA -1 C09AA	19.1	8.678	2.339
Female	60+	C09AA -1 G03CA	17.8	7.687	2.280
Female	60+	C08CA C09AA -1 C08CA -1 C09AA	17.8	7.144	2.217
Female	60+	C09AA -1 C08CA -1 C08CA -1 C09AA -1 C09AA	17.8	6.491	2.139

4 Conclusion and Discussion

We have presented a temporal sequence mining framework for rare events and successfully identified interesting associations and sequential patterns of drug exposure which leads to a high risk of certain severe adverse reactions. Our intent is to generate hypotheses identifying potentially interesting patterns, while we realise that further validation and examination are necessitated. We also note that matching of patients has not included matching for time. It is proposed that case matching for time will be a refinement which could improve the approach to reduce the potential for false positives. In estimating risk ratios, it is not clear how to calculate confidence intervals in this case. Confidence intervals are important in identifying the degree of uncertainty in the estimates, and further work is required to investigate this deficiency. Besides adverse drug reactions, our approach may also be applied to a wide range of temporal data mining domains where rare events are of primary interest.

References

1. P. Armitage, G. Berry, and J. N. S. Matthews. *Statistical Methods in Medical Research*. Blackwell Science Inc, 4 edition, 2002.
2. S. D. Bay and M. J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 302–306, 1999.
3. G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52, San Diego, August 1999.
4. L. Gu, J. Li, H. He, G. Williams, S. Hawkins, and C. Kelman. Association rule discovery with unbalanced class. In *Proceedings of the 16th Australian Joint Conference on Artificial Intelligence (AI03), Lecture Notes in Artificial Intelligence*, Perth, Western Australia, December 2003.
5. M. Seno and G. Karypis. SLPMiner: An algorithm for finding frequent sequential patterns using length decreasing support constraint. In *Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM)*, pages 418–425, Maebashi City, Japan, Dec 2002. IEEE.
6. G. I. Webb. Efficient search for association rules. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 99–107, 2000.
7. G. Williams, D. Vickers, R. Baxter, S. Hawkins, C. Kelman, R. Solon, H. He, and L. Gu. The Queensland Linked Data Set. Technical Report CMIS 02/21, CSIRO Mathematical and Information Sciences, Canberra, 2002.
8. W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. WSARE: What's strange about recent events? *Journal of Urban Health*, 80(2):i66–i75, 2003.