

A Sequence Building Approach to Pattern Discovery in Medical Data

Jorge C. G. Ramirez^{1,2}, Lynn L. Peterson¹, and Dolores M. Peterson²

¹Department of Computer Science & Engineering, University of Texas at Arlington,
PO Box 19015, Arlington, TX 76019-0015
{ramirez, peterson}@cse.uta.edu

²HIV Clinical Research Group, Division of General Internal Medicine, University of Texas Southwestern Medical Center
5323 Harry Hines Boulevard, Dallas, TX 75235-9103
dpeter@mednet.swmed.edu

Abstract

The goal of the research being reported is the discovery of useful concepts in temporal medical databases. In this paper, we present a sequence building approach, based on the Generalized Sequential Patterns (GSP) Algorithm (Srikant and Agrawal 1996), to discover temporal patterns in this type of data. We show that this pattern discovery is possible by normalizing continuous features in a way that allows meaningful comparisons between patients.

Introduction

There has been a recent proliferation of literature on knowledge discovery in databases (KDD) and data mining (DM). Most data mining techniques require that the data be in some standard form. However, many databases, and in particular medical databases, have features that make them different from the type of data collections used with most KDD/DM methodologies. Specifically, the data may be any combination of binary, numeric, symbolic and text data. In addition, the data can be temporal, with different significance attached to the temporal aspect, depending on the specific data type. Furthermore, the data fields themselves generally are not the same at each collection point. The purpose of our research is to identify methodologies that will be useful in the discovery of patterns in such data. The current goal is to treat clinical events as sets, and to discover patterns in sequences of these event sets. This paper discusses the first steps in this direction, using an approach based on the Generalized Sequential Patterns (GSP) Algorithm [Srikant & Agrawal 1996].

We are interested in discovering patterns in data that span the course of disease. Given a database that contains clinical data for patients diagnosed with a specific catastrophic or chronic illness, we are interested in discovering patterns in that data that show that groups of patients had similar experiences during the course of the disease. The

motivations for such research are many. With advances in medical technology have come many methods for treating such illnesses. Analysis of the course of such diseases is beneficial from multiple points of view, including enhancement of provision of care, prognosis, monitoring, outcomes research, cost/benefit analysis, and quality assurance. This type of research is also beneficial for development of techniques of pattern discovery for other data collections that have similar characteristics.

Approach

From the characteristics of the data alone, it is easy to see that the long-term goal of this research has many complicating factors. Using the top-down approach, we are subdividing the problem and addressing individual issues one at a time. The first is the issue of standard form for continuous variables. Particularly in medical patients with catastrophic or chronic illnesses, it is not possible to use general population norms to judge the current state of the patient's progress through the disease process. That comparison will almost always indicate that the patient is not well, without regard to how he or she is doing with respect to other patients with the same affliction. Each patient will settle into his or her own individual norm for the various clinically-measurable data.

Domain

We are using the Jonathan Jockush Human Immunodeficiency Virus (HIV) Clinical Research Database (JHIVCRDB) developed and maintained at the University of Texas Southwestern Medical Center. This database contains records for over 8,500 HIV-positive patients dating back to 1987. In an effort to generate sequences of sufficient length and content from which to discover interesting concepts, we use only patients that have been followed for at least 48 months and have at least 30 dates on which

¹Copyright © 1998, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Table 1. Results of Normalization Procedure Applied to White Blood Cell (WBC) Count

a. HIV patient, clinically deteriorated		b. Norms, general population		c. HIV patient, relatively normal	
Normalized Value	Actual WBC Range	Normalized Value	Actual WBC Range	Normalized Value	Actual WBC Range
-4	up to 1.3	-4	up to 2.9	-4	up to 2.5
-3	1.4 - 2.0	-3	3.0 - 4.0	-3	2.6 - 3.7
-2	2.1 - 2.7	-2	4.1 - 5.1	-2	3.8 - 5.0
-1	2.8 - 3.4	-1	5.2 - 6.3	-1	5.1 - 6.2
0	3.5 - 5.2	0	6.4 - 8.6	0	6.3 - 8.6
+1	5.3 - 6.4	+1	8.7 - 9.8	+1	8.7 - 9.8
+2	6.5 - 7.7	+2	9.9 - 10.9	+2	9.9 - 11.1
+3	7.8 - 8.9	+3	11.0 - 12.0	+3	11.2 - 12.3
+4	9.0 and up	+4	12.1 and up	+4	12.4 and up

laboratory data was collected. There are approximately 1100 such patients in the database. For our early experiments we used a random subset of this group of patients.

Normalization of Laboratory Results

White Blood Cell Count. Specifically, we started by examining white blood cell (WBC) count in patients with HIV, which may or may not have progressed to Acquired Immune Deficiency Syndrome (AIDS). Standardizing values between event types is not the issue; standardizing values between patients is. What is normal for one patient can range from normal to very poor for another patient. Using standard statistical techniques, we developed a methodology for normalizing WBC counts to a range of -4 to +4 where 0 is normal, and both -4 and +4 are indicative of severe illness. The choice of -4 to +4 was made to represent the approximate number of standard deviations away from the “normal” average. The methodology is based on norms for the general population with adjustments made for the fact that HIV compromised patients tend to have lower than normal values. However, the methodology takes into account the fact that the normal value for any given patient may be different, or not, compared to the general population.

A comparison of results is shown in Table 1. The general population norm is shown in Table 1b. This can be compared to the results obtained by using our methodology on a clinically deteriorated patient in Table 1a, and a healthy HIV patient in Table 1c. Notice that “normal” (i.e., 0) for our clinically deteriorated patient is 3.5 to 5.2, which is poor (i.e., ranging from -3 to -2) compared to the general population; however “normal” for our relatively normal patient is 6.3 to 8.6 which is almost exactly the same as the range for the general population.

Generalization. WBC is a member of a class of clinical data with the following characteristics: 1) some median value is normal and therefore good; 2) a significant increase in value is a clinical indication of a disease process in action; and 3) a significant decrease in value is also indicative of a problem with the patient's health status. Given the number of features involved, it would not be expedient to do such a detailed analysis on every individual variable. We chose to analyze all variables with the same characteristics as a group. Taking the specific numbers that were used in the WBC equations and relating them back to how they were developed, i.e., by parameterizing the equations, generic equations were created to attempt analysis on other data of the same class. We applied these parameterized equations to a few chosen features that are members of the same class of clinical data as WBC: Hematocrit (HCT), Platelets (PLT), CD-4 percent (CD4P), CD-4 absolute (CD4A), and Lymphocytes (LMPH) with results similar to those for WBC. When the results were reviewed by the clinicians in the HIV Clinical Research Group, they were impressed that the results matched up with their clinical experiences.

Event Set Sequences

The basic approach for organizing our data involves the use of sets of data which are grouped by date and include treatments and measures of health status. The algorithm iteratively searches for sequences of event sets common to multiple patients. Temporal windows are used to group together related events that occur close to each other but not on the same date. Formalizing this algorithm uses the notion of Event Set Sequences (ESSs) of patient data, of which the formal definition is given below.

$$\begin{aligned}
&< \{ \text{HCT } 0 \quad \text{PLT } 0 \quad \text{CD4P } 0 \quad \text{CD4A } 0 \quad \text{LMPH } -1 \} \{ \text{WBC } 0 \quad \text{HCT } 0 \} \\
&\quad \{ \text{PLT } 0 \} \{ \text{WBC } 0 \quad \text{HCT } 0 \quad \text{PLT } 0 \quad \text{CD4P } 0 \quad \text{CD4A } 0 \quad \text{LMPH } 0 \} \\
&\quad \{ \text{WBC } 0 \} \{ \text{WBC } 0 \quad \text{HCT } 0 \quad \text{PLT } 0 \} \{ \text{HCT } 0 \} \{ \text{WBC } 0 \} \{ \text{WBC } 0 \} >
\end{aligned}$$

Figure 1. Discovered Pattern. Uses 6 event types, 33% support threshold and 90 days maximum gap. Contains 9 clinical event sets, potentially spanning over 2 years.

Definitions

Let $O = \{O_1, \dots, O_j\}$ be the set of *occurrence events*, where O_i has duration $d_i \in DO_i$, the set of possible durations of O_i . Let $V = \{V_1, \dots, V_k\}$ be the set of *value events*, where V_i has domain DV_i , the set of possible values of V_i . Let $e = (id, t, E, v)$ be an *event*, a four-tuple where id is the patient, t is the time of e , $E \in (O \cup V)$, and if $E \in O$, specifically O_i , then $v \in DO_i$, otherwise $E \in V$, specifically V_i , and $v \in DV_i$. Let $ES = \{e_1, \dots, e_m\}$ be an *event set*, a non-empty set of events, where $id_1 = \dots = id_m$, and $t_1 = \dots = t_m$. Let $ID_i (= id_1 = \dots = id_m)$ be the actor for ES_i , and $T_i (= t_1 = \dots = t_m)$ be the time of ES_i . Finally, let $ESS = \langle ES_1, \dots, ES_n \rangle$ be an *event set sequence*, an ordered list of event sets where $ID_1 = \dots = ID_n$, and $T_1 < \dots < T_n$.

In our domain, O contains pharmacy events (i.e., the dispensing of medications) and diagnosis events (e.g., a pneumonia), and V includes charge events (e.g., for clinic visits or hospital stays) and laboratory-test-result events (e.g., white blood cell count). Time is measured in days; therefore, each Event Set (ES) is a collection of those events that occur on a specific day. Finally, an Event Set Sequence (ESS) is an ordered list of Event Sets for a given patient.

We seek to discover patterns in ESSs such that there is a similarity among patients in at least a sub-sequence of the sets of events. We say that a pattern is discovered for a group of ESSs if we can find some sequence $S_i = \langle s_1, \dots, s_q \rangle$ for each ESS_i , where s_i denotes the subscript of an ES, $1 \leq s_1 < \dots < s_q \leq n$, q is equal for all ESS_i in the group, and each ES of the corresponding sequence “matches” across all ESS_i . The definition of “match” is therefore crucial to the discovery process. While this definition is continually being reviewed, in the current model we define $ES' \subseteq ES$, where all $e \mid e \in ES'$ match exactly across all ESSs; however, the definition does not require all $e \in ES$ be present in ES' . Further, we use a windowing technique, as in the GSP algorithm, that allows for events that do not actually occur on the same day, but occur within the specified amount of time, to be included as part of the same ES or ES' .

Pattern Discovery

We developed a variation of the GSP Algorithm to discover patterns in our domain. Though the basic algorithm is the same, the details of implementation are different. The GSP algorithm was designed to work on sequences of events that either occurred or did not, where the occurrence or lack thereof was significant to the patterns discovered. None of the events had attributes.

The differences in domains leads to several significant observations, particularly for future work. In our domain, the occurrence or lack thereof does not have any specific significance. The events themselves have attributes, especially when viewed from the event set perspective. Finally, the sheer numbers of events being dealt with strains an algorithm that was designed to discover patterns at an individual event level.

Results

Initially, we ran our modified GSP Algorithm on a random sample of six patients who had all been followed for 60-65 months and had 30-39 dates during that time period in which clinical data was collected. As it turned out, of the original six patients, two of them were particularly “well” over the time period and therefore, the longest sequences that were discovered were for groups of patients that were well. However, we believe that once we pull more patient data from the available 1000+ patients that have been followed for a minimum of 48 months, with a minimum of 30 clinical event days, we will ultimately discover patterns for groups of patients with varying clinical patterns. For example, one experiment (where 6 event types (or features), a support threshold of 33%, and a maximum gap of 90 days between clinical event sets was used) resulted in the pattern shown in Figure 1 being discovered.

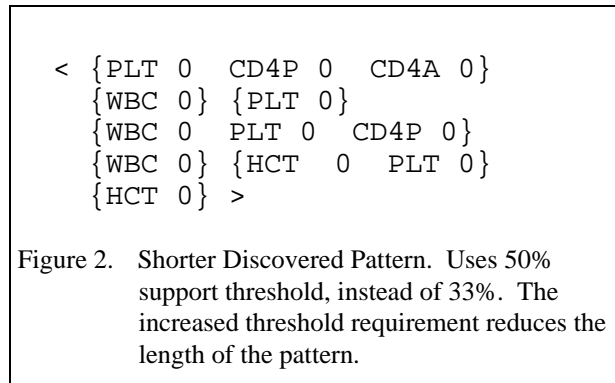
This pattern represents 9 clinical event sets that are supported by the database, where the first one matches on five of the six features, the second matches on two of the features, etc. Notice that the fourth event set matches on all six features and that the entire pattern represents a “well” patient over the period since nearly all of the standardized values are 0. This result does not mean that the patients that supported this pattern did not have periods when they were not as well, but that if they did, they did not have it in a sequence that paralleled the other patients that supported this particular pattern. Given the parameters

used for this experiment, this particular pattern potentially spans over 2+ years.

Effects of Varying Parameters on Pattern Length

We have varied many parameters of the algorithm to get a feel for which ones have the most effect on the length and interestingness of patterns discovered, as well as the amount of time taken to discover those patterns. Gut instinct is borne out by the numbers; as the parameters allow for longer sequences to be discovered, the time for the discovery increases. The patterns are generally longer when more features are available to choose from, when the support threshold is lower and when the maximum gap is higher. Our initial experiments used the GSP Algorithm and a single feature, WBC. However, we have now modified the algorithm to incorporate any number of standardized features.

Support. Support is the percentage of all patients whose ESS actually contains a given pattern. Srikant and Agrawal employed a 40% support threshold in their domain. Since we started with 6 patients, we used 50% and 33% in our initial experiments. As expected, the 33% threshold consistently resulted in longer patterns. When the same experiment as shown in Figure 1 was run with a 50% threshold, as opposed to the 33% threshold, it resulted in the shorter pattern shown in Figure 2.



Maximum Gap. Since even an HIV patient who is doing relatively well sees the doctor at least every 2-3 months, we started with a maximum gap of 90 days; however, there was a significant difference in pattern length when we increased that gap to 120 days. If a patient has a gap in his or her ESS longer than the maximum gap, then it has the effect of partitioning that patient's ESS into subsets. The likelihood of a longer pattern being supported by that patient is therefore decreased. However, if there are no gaps in the ESS longer than the maximum gap, then that patient is more likely to support longer patterns.

Computation Time

Inversely, the amount of time taken to discover patterns is less when there are fewer features being used, when more support is required from the database, and when the maximum gap allowed between clinical events is smaller. One result, though not surprising, is that we have experienced an exponential explosion in time when we decrease the support required and increase the maximum gap. We have made a few modifications and improved times over our initial experiments; however, goal directing techniques will need to be incorporated to prevent exhaustive searches that take an unsatisfactorily long time.

Table 2 shows a comparison of the computation time in terms of CPU seconds and nodes searched for

a. <u>Threshold = 50%</u>				
Time(CPU Secs)/ Nodes Searched(100,000s)				
# Features	<u>6 Patients</u>	<u>12 Patients</u>	<u>24 Patients</u>	
3	9 / 11	4 / 6	352 / 262	
4	15 / 17	5 / 7	371 / 278	
5	32 / 37	6 / 8	448 / 336	
6		7 / 9	638 / 462	

b. <u>Threshold = 33%</u>				
Time(CPU Secs)/ Nodes Searched(100,000s)				
# Features	<u>6 Patients</u>	<u>12 Patients</u>	<u>24 Patients</u>	
3	102 / 66	64 / 77	9341 / 6102	
4	345 / 152	72 / 91	Not Available	
5	3858 / 914	93 / 118	Not Available	
6		112 / 144	Not Available	

experiments where the maximum gap is 90 days, the threshold is 33% or 50%, the number of event types included varies from 3 to 6, and the number of patients is 6, 12 or 24. There is a linear-order increase in computation time as more event types (features) are added. The exception is the 6 patient group when the addition of the 5th feature (CD4 Absolute) caused a dramatic increase. We believe this was an aberration due to the small sample size, and the fact that this particular feature is collected frequently in HIV/AIDS patients. Compared to the 6 patient group, computation time dipped for the 12 patient group, then increased alarmingly for the 24 patient group. Again we believe the dip was an aberration, this time attributable to the specific group of 12 patients involved.

Interestingness

Goal directing techniques will also need to be used to prevent discovery of long patterns of relatively low interest, though the issue of interestingness is yet to be fully explored. An example of a pattern that might be more interesting is shown in Figure 3:

$$\langle \{WBC\ 0\} \{HCT\ -4\} \{PLT\ 0\} \\ \{WBC\ 4\} \{WBC\ 1\} \rangle$$

Figure 3. Potentially More Interesting Pattern. This pattern alternates between periods of wellness and illness.

The pattern seems to alternate between periods of wellness and periods of illness. This pattern was discarded after there was not support in the database to carry it out further. This fact leads us to believe that in order to discover patterns with some interestingness to them, we will need to drop the support threshold, since we are in fact interested in groups, which are not necessarily large, with similar course of disease.

Related Work

(Mannila, Toivonen and Verkamo 1995) address the temporal aspect of discovery, locating frequently occurring episodes (i.e., combinations of events with a partially specified order) from a long sequence of events. (Mannila and Toivonen 1996) extend the technique to include specification of order of events and discovery of general episodes. Padmanabhan and Tuzhilin also extend (Mannila, Toivonen, and Verkamo 1995), noting that the work applies to sequences, introduce temporal logic as an appropriate formalism for expressing temporal patterns in categorical data, and then use it as a means to discover patterns in temporal databases. Of course, (Srikant and Agrawal 1996) present the GSP algorithm that incorporates time constraints to specify maximum and/or minimum time gaps between adjacent elements of a sequential pattern. Further, they generalize the definition of the patterns to incorporate taxonomies in the data.

While all of this work is related, and certainly we use the basic approach of the GSP algorithm, the domains are much more restricted and lack the complexities of the medical data domain. Specifically, the example domains used had specific meaning to the occurrence or lack of occurrence of an event. This is not true in our domain. The lack of occurrence of a given event during a given event set does not necessarily have specific meaning, other than the fact that that event was not recorded at that time. Further, the occurrence of a given event may not have specific meaning. While some of the domains had attributes associated with the events, support

was determined based on exact match, which cannot work in the medical data domain.

Conclusions and Future Work

Though the current results are somewhat interesting, the modified GSP Algorithm does not actually accomplish our long-term goal. As currently implemented, the algorithm requires that all features be present to support the pattern; however, since the set of recorded data that exists for each patient not only varies greatly between patients, but also varies between collection dates for any given patient, our next goal is further modifications to account for missing data. Our event set strategy should help to address this issue and reduce the computational complexity as well. Currently every individual event is analyzed for support of a pattern. By comparing event sets at the set level, we believe that computational complexity will decrease significantly. Other future goals include the aforementioned investigation into interestingness of discovered patterns, which will include dropping the support threshold, adding goal directing, and the incorporating temporal aspects of features not included in this work (e.g., the duration of a diagnosis or a treatment).

Acknowledgments

Jorge Ramirez is supported in part by NSF grant GER-9355110.

References

- Mannila, H., H. Toivonen, and A.I. Verkamo. 1995. Discovering Frequent Episodes in Sequences. Proceedings, 1st International Conference on Knowledge Discovery and Data Mining, p 210.
- Mannila, H. and H. Toivonen. 1996. Discovering Generalized Episodes Using Minimal Occurrences. Proceedings, 2nd International Conference on Knowledge Discovery and Data Mining, p 146.
- Padmanabhan, B. and A. Tuzhilin. 1996. Pattern Discovery in Temporal Databases: A Temporal Logic Approach. Proceedings, 2nd International Conference on Knowledge Discovery and Data Mining, p 351.
- Srikant, R. and R. Agrawal. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. Proceedings, 5th International Conference on Extending Database Technology, p. 3.