Temporal Discovery Workbench: a Case Study with ICU Patient Datasets

Derek Sleeman^{1,2}
¹Dept. of Computing Sc.,
University of Aberdeen,
Aberdeen, AB24 3UE
d.sleeman@abdn.ac.uk

Laura Moss^{1,2}
²School of Medicine,
University of Glasgow,
Glasgow, G31 2ER
laura.moss@glasgow.ac.uk

John Kinsella
School of Medicine,
University of Glasgow,
Glasgow, G31 2ER
john.kinsella@glasgow.ac.uk

Temporal datasets are now often collected and curated in industry, scientific labs, and healthcare. A considerable amount of work has been done to analyze these for trends, inconsistencies and to make use of them for prediction. In an earlier study we discussed datasets for 15 or so Intensive Care Unit patients with clinicians, and asked them if they could detect when a particular harmful event (a myocardial infarction) had occurred. After 3 lengthy knowledge capture sessions we formulated a complex model to identify myocardial damage which we then implemented and tested against a test dataset; a detection rate of approximately 80% was achieved. (Specifically the model suggests that several events generally occur in a temporal sequence before the event-to-be-predicted occurs.) This work reports the design of a Temporal Discovery Workbench (TDWB) to address this class of tasks and has reproduced the results of the initial model acquired from the experts. Further we have now run TDWB's pattern discovery module with a range of settings to see if further clinically useful patterns are reported. Initial results are encouraging.

Keywords: Modelling of Expertise, Temporal Datasets, Event Prediction, Workbench, Intensive Care Unit, Myocardial Damage.

1. INTRODUCTION

Temporal datasets are now often collected and curated in industry, scientific labs, and healthcare. For example, in medicine, it is now common to collect large amounts of data for patients, and often a sizable number of time-points are recorded during the course of a treatment. This is particularly true for patients in Intensive Care Units (ICUs) which often collect 50 or so descriptors at least hourly, and for such patients to stay in these wards for many days. These datasets are rich repositories which have been used both to determine whether descriptors are correlated, when it should be possible to either demonstrate a deterministic or non-deterministic (e.g. *statistical*) relationship between the descriptors. Such relationships have been expressed in a variety of forms including polynomial functions and rules of the form:

if A and B and C and not(D) then...

Generally, once the above relationships have been formulated it is then possible to predict the values of a particular descriptor at some point in time, based on the other descriptors either in the same time-slot or earlier ones.

1.1 Context

Earlier, we reported a study where we asked experts to discuss conditions under which Myocardial Damage can occur in ICU patients and then compared the results of their model with a test dataset (Sleeman et al, 2011). Here's a summary from that report: "Myocardial damage is known to occur relatively frequently, and although it is not often fatal it results in the patient staying in the ICU for significantly longer. Thus it is important for clinicians to detect these events. Confirmation of myocardial damage is by a biomarker (troponin), but these tests are only done at fixed time-points. Consequently it is desirable for doctors, and support systems, to detect myocardial damage from the standard descriptors collected for ICU patients. We have undertaken a study with several ICU consultants to determine the conditions which generally precede a myocardial-damaging event. In fact, these knowledge acquisition sessions produced a complex model which we have realized as 2 interacting modules. Subsequently, we compared this model's predictions against the original datasets; the model when run against the test dataset resulted in a relatively high True Positive (TP) rate (75.8%)" (Sleeman et al, 2011).

This was a very encouraging result. However the model articulated by the experts as indicated above was relatively complex. The following is a slightly simplified summary of the conditions under which the experts believe Myocardial Damage (MD) occurs:

- **MD** is confirmed when a cardiovascular derangement (CVD) sequence is followed by a raised troponin value within [1-72] time-slots.
- a **CVD sequence** is said to occur when CVD (cardio-vascular derangement) events occur in at least 3 out of 5 adjacent time-periods
- a **CVD event** is said to be either:
 - a very extreme value for any of the following patient descriptors: SpO2 (Oxygen Concentration in the patient's blood), HR (Heart Rate) or MAP (Mean Arterial Pressure i.e., the patient's Blood Pressure)¹. Note there are 5 possibilities as HR and MAP can have both extremely low and extremely high values.
 - a combination of 2 of the above descriptors with extreme values (Giving 8 combinations)¹
 - a combination of 3 of the above descriptors with considerably abnormal values (but less severe than "extreme").
 - extremely high levels of FiO2 (inspired oxygen) or a rapid increase in FiO2 between several time-points.

As a result of this study we decided to develop the Temporal Discovery Workbench (TDWB) to see whether given background information about the domain, and the same temporal sequences as the experts analyzed, firstly the TDWB would be able to replicate the (complex) model articulated by the experts and secondly whether it would be able to suggest some alternative, possibly simpler, models/patterns.

We are addressing the general scenario in which an unusual event, E, happens at time-point, T, and we aim to predict this event by analyzing trends and absolute values in the several descriptors recorded in the time-period prior to E. To help this analysis, it is likely we will also have datasets involving the same descriptors in which the event, E, **does not** occur.

1.2 Advantages of Workbenches

As mentioned, we have decided to implement a Workbench (Temporal Discovery Workbench - TDWB) as we believe this gives us a great deal of flexibility. Specifically although we have a clear idea of the project's overall objectives we do not know in advance the range of relevant applications we might encounter and thus we do not know the detailed nature of the analyses which domain experts might wish to carry out on their datasets. Workbenches (WBs) generally present their user with options at each step in the analysis and allow the analyst (sometimes with guidance) to decide the data display mode or analysis package to be used and with what descriptors. It is essential that WBs provide user friendly interfaces, and they are modular in construction, so that functionality not envisaged at the initial design can be subsequently added if needed. (In later sections we discuss the outline implementation of the TDWB.)

1.3 Overview of the Paper

Section 2 gives brief literature reviews of the analysis of temporal datasets and the Apriori algorithm. Section 3 outlines the functionality of the TDWB (Temporal Discovery Workbench). Section 4 reports the results of analyzing the Glasgow MD patient dataset with TDWB. Section 5 discusses plans for further work.

2. LITERATURE REVIEW²

As noted in the introduction, temporal datasets are now regularly collected by many companies and institutions and there has been considerable interest in analyzing these datasets for example to detect inconsistencies, trends, recurrent patterns etc. Combi et al (2010) gives a good overview of the use of temporal Information Systems in Medicine.

An important development in data mining has been the ability to establish that a descriptor is associated with one or more other domain descriptors; Agrawal & Skikant (1994) developed the very efficient Apriori algorithm to detect such patterns. Laxman & Sastry (2006) have subsequently developed this approach so that it is able to detect patterns in temporal datasets. And of course if these temporal patterns are applied uni-directionally then causality is suggested.

TDWB attempts to infer association patterns between the descriptors in the (temporal) domain it is analyzing.

¹ For a detailed discussion of the descriptors recorded regularly for each ICU patient, and a scale used to describe the patient's status, see (Sleeman et al, 2009).

² Due to space limitations, a more general review of modelling of medical domains has not been included; see for example (Sleeman et al, 2011).

The general form of the patterns/rules which it infers is:

*IF A@T+0³ and B@T+1 occur
THEN expect E[T+2, T+50].*

As the Apriori algorithm – and more particularly the temporal extension – are central to TDWB, this approach is outlined in some detail here. The Apriori approach identifies frequently occurring individual items in a dataset, and extends these to larger descriptions as long as those descriptions appear sufficiently often in the dataset. (The analyst is usually required to set the minimum level of support for the evolving patterns). The central insight behind this algorithm is that whilst there may be many objects with a single descriptor in a database, this number drops off once the number of descriptors in a pattern is increased. And as a result the search space is usually relatively limited even for large datasets. Once identified, these descriptors can be used subsequently as association rules. We now discuss the extension to temporal datasets as discussed by Laxman & Sastry (2006). Suppose each of the 5 lines given below corresponds to the objects purchased by Customers 1....5:

1. <(A B) (A C D) (B E)>
2. <(D) (A B E)>
3. <(A) (B D)(A B E F) (G H)>
4. <(A) (F)>
5. <(A D) (B E G H) (F)>

More specifically, the information between <..> represents the goods bought by Customer (i) on the various occasions when they visited the store. So for instance, Customer (2) went shopping on 2 occasions: the first time they purchased D and the second time they purchased 3 items (A B E). Note the order of the shopping visits is significant, but the items, bought at any one visit, are not significant. The temporal Apriori algorithm, given a support level of 2 would extract the following pattern <(D) (G H)> which implies, in the above records, there are 2 shoppers who on an earlier visit to the store purchased "D" and on a subsequent visit purchased "G" and "H".

3. OVERVIEW OF THE TEMPORAL DISCOVERY WORKBENCH

As noted earlier, TDWB addresses the general scenario in which an unusual event, E, happens at time-point, T, and aims to predict this event by analyzing trends and absolute values in the several descriptors recorded in the time-period prior to E.

Temporal datasets are presented to TDWB as CSV files, and must contain a column called Time-point (containing data of the following form: [DD:MM:YYYY; hh:mm:ss], and a column called "Special Event" which can only contain the strings: Positive, Negative or blank. Additionally, the file can contain as many other column headings as required by the domain. So in the case of the ICU domain this is likely to include: variables such as HR, Mean (or MAP), FiO2, SpO2, together with drugs information. Each column is typed to help TDWB spot data errors; currently only the following data types are accepted: "Timepoint", "Int" (integer), "Real" and "String". The data associated with a particular time point is held as a separate record; each record is terminated by a New Line; and files are terminated by a special terminator. The workbench has essentially 3 phases, namely: Data files (Input), Data Analysis, and Pattern Matching and Discovery which are discussed below.

Figure 1: This screen allows the analyst to select descriptors to be used and to set up the several ranges for the SpO2 Descriptor.

"Data files" loads patient (CSV) files, performs various checks on the dataset (including: type-checking of elements, that temporal records are correctly ordered, check length of gaps between time-points), provides options for extrapolation of missing time-points etc; allows the analyst to select from all the descriptors in the CSV file which should be included in the current analysis; and set ranges for the selected descriptors. For example (see Figure 1⁴), the expert decided that SpO2 should have the following 5 ranges: L4 (Low-4), L3, L2, L1 and N (Normal). Multiple patient datasets can be

³ Where A@T+0 refers to observation A which occurs at time point 0

⁴ Six full-sized screenshots are available for this paper; please email d.sleeman@abdn.ac.uk for a set.

loaded. There are also facilities to display the datasets in different formats: raw/original, cleaned-up (i.e., when extra and missing elements/time-points are dealt with), continuous data with a predefined set of ranges for each descriptor, and discrete where the names of the ranges are displayed. Once these processes have been successfully completed, the analyst is given the opportunity to save this information to a project file so that the "set up" work does not need to be done again.

"Data Analyses" is not very highly developed as yet; for details of its current features see the User Manual, (Sleeman & Blasco, 2012).

The "Pattern Matching and Pattern Discovery" module provides the most extensive set of facilities. In this summary, due to space limitations, we describe just the relevant subset of TDWB's functionality. The pattern creating modules allow the analyst to select the segments, descriptors, and descriptor ranges that are to be used in a run of the pattern generation algorithm. Additionally the analyst is able to decide whether elementary or composite elements are to be the building "blocks" for the temporal patterns (see Table 1 for these definitions), the minimum and maximum number of temporal elements to be included in each pattern, and the maximum number of gaps to be included in each temporal pattern.

Table 1: Elementary, Composite and Temporal Patterns Definitions

Elementary Patterns	extend over a single time-period, and are of the following form: SpO2[L4]@T+0 i.e., they contain a descriptor-value pair and specify a time-point (i.e., T+0)
Composite Patterns	are ones that involve 2 or more elementary patterns associated with a single time-point, e.g., SpO2[L4]@T+0, HR[H3]@T+0
Temporal Patterns	are defined as patterns which extend over one or more time-points; further they can contain Gaps i.e., time-points where none of the "active" descriptor-value pairs occur. An example of a temporal pattern is: SpO2[L4]@T+0, GAP, HR[H3]@T+2. Additionally, the elementary and composite patterns mentioned above are also valid temporal patterns.

Another very important parameter used by the Pattern Generation algorithm is the "Positive Threshold" parameter which specifies the number of PSEs (i.e., segments that have as their last element a Positive Special Event marker) which should be matched by any pattern generated. Ideally as a result of the pattern generation process we will end up with a small number of patterns which cover all

the PSEs and none of the NSEs (i.e., segments that have as their last element a Negative Special Event marker); in most real-world situations where data is noisy this is unlikely to be the case. Early on we made a design decision to make the processes of Pattern Generation and the determination of pattern "Coverage" distinct modules. There are several reasons for doing that: firstly, the processes are then much more transparent to the domain expert, and secondly if one wishes later to implement, e.g., a more sophisticated coverage algorithm one needs only to add this new algorithm to the coverage module, and one does not need to modify the Pattern Generation module. This level of modularity is consistent with the workbench philosophy we articulated earlier. So in all "core" studies to date we have set this parameter to 1, so that the Pattern Generation modules report the various patterns which are found for each of the PSEs, and the analyst (with some support from the WB) then selects in the Coverage module a set of patterns which satisfy, as best as it can, the particular trade off criteria they wish to apply between covering all PSEs and no NSE.

4. TDWB: CASE STUDY OF MYOCARDIAL DAMAGE ANALYSIS / PREDICTION

4.1 Studies Description

In section 1, we summarized the model which we formulated as a result of several knowledge acquisition sessions with 2 domain experts, and we reported the results which that model achieved when it was applied to the test set of 34 patients (a relatively high True Positive (TP) rate (75.8%), (Sleeman et al, 2011). What needs to be stressed here is that this model reports an association between the CVD sequences and a raised troponin value - that is a positive correlation is recorded if the CVD sequence occurs either before or after the raised troponin provided these events are within the defined time window of 72 hours. However, being able to predict that a CVD sequence is always/frequently followed by a raised troponin value is of course much more useful clinically.

We have since run this expert model to determine how effective their model is at prediction [Moss et al, 2012]; prediction is the focus of the analyses we have now undertaken with TDWB. Also we should point out that we are reporting the results of fewer PSE & NSE segments as TDWB's loading module found a number of inconsistencies in some patient datasets which had previously not been detected. (For example a time gap between 2 records of over a year.)

The studies reported here have been run with a total (across all patients) of 23 PSEs & 15 NSEs (segments). We summarize the various descriptors used with each of the studies in Table 2. *Note: The*

patterns effectively predict that a raised troponin will be detected within 72 hours of the CVD described by the several temporal patterns produced.

4.2 Comments on Patterns produced by each of the Studies.

The first column in Table 3 gives the study number i.e., Study-1 to Study-24. The "All" column reports the number of patterns created by the Apriori algorithm for that study (with the combination of descriptors specified in detail above). Because this is often a sizable number of patterns we have implemented a facility by which the Coverage module is able to select for each PSE the N highest ranked patterns. So "ALL1" corresponds to the patterns selected when the algorithm is retaining just the top ranked pattern for each PSE or all such patterns if a set of patterns are given equivalent ranking.

The ranking of patterns is done by assigning a positive value for each PSE matched by a pattern, and a negative value corresponding to each NSE matched. So in the case of this study, the domain expert suggests a +5 and -2 respectively; note these values are parameters and can be changed for each analysis. As you can see from looking over the figures this filter is quite effective at reducing the number of patterns to be considered (in the case of Study-2 the reduction is from 339 to 43).

The 4th column provides the usual metrics (True Positive (TP)/False Negative (FN)/True Negative (TN)/False Positive (FP)) for both the All & All1 sets of descriptors; it can be easily shown that they should be identical. As mentioned earlier the role of the Coverage module is to help the analyst/domain expert select patterns; one common objective is to cover as many of the PSEs as possible, and as few NSEs as possible (See figure 2.) This, in general, is clearly a complex optimization process, so for the moment we report the results for "All1" and for another straightforward, but effective, strategy where any pattern which matches a NSE is removed; this is the NoNSE strategy reported in the table.

Table 2: Studies Description

Study-1:
<i>This uses TDWB to rerun the "manual" model obtain from the clinicians over a (smaller) dataset. Here the base patterns used are: SpO2[L4], HR[L4], HR[H4], MAP[L4], or MAP[H4]; 2 of the above descriptors at level-3 (i.e., L3 or H3); 3 of the above descriptors at level-2 (i.e., L2 or H2); and FiO2[H4]⁵. Also following the initial model, each reported pattern must have one of the above sub-patterns occurring at 3 out of 5 time-points (that is the model allows up to 2 gaps in each of the patterns); and this must be followed within 72 hours by a PSE (i.e., a raised troponin value)</i>
Study-2:
<i>The remaining studies have used the (degenerate)⁶ Apriori algorithm to create temporal patterns; and in all remaining studies we have just used 3 descriptors: SpO2, HR, and MAP; in all these studies the algorithm could, if supported by the data, suggest composite patterns. In the case of study-2 we excluded from the analysis descriptor-ranges which were N (Normal) & those at level-1 (i.e., L1 & H1). Here we specify that the minimum number of elements (elementary or composite patterns) must be 1, and the maximum number of elements must be 3. Further we specified that temporal patterns can include up to 2 gaps, so the length of the temporal patterns produced are between 1 and 5 units</i>
Study-3:
<i>All the parameters are the same for those in Study-2 except that descriptor-ranges at level-2 (i.e., L2 and H2) are also now excluded</i>
Study-4:
<i>All the parameters are the same for those in Study-3 except that descriptor-ranges at level-3 (i.e., L3 and H3) are also now excluded</i>
Study-22:
<i>All the parameters are the same as for Study-2 except that now we specify that the minimum & maximum number of elements in a temporal pattern must be 3. But up to 2 gaps are still possible, and so the temporal patterns produced here can be between 3 and 5 time-units in length. (Whereas those produced in Study-2 can between 1 and 5 units in length.)</i>
Study-23:
<i>All the parameters are the same as for Study-3 except that now we specify that the min. & max. number of elements in a temporal pattern must be 3</i>
Study-24:
<i>All the parameters are the same as for Study-4 except that now we specify that the min. & max. number of elements in a temporal pattern must be 3</i>

⁵ This is a slight simplification of FiO2's role in the model as specified by the domain experts.

⁶ The degenerate Apriori algorithm is a combinatorially simpler version which is applicable only when one is

seeking patterns which cover a single PSE (Positive Segment)

Table 3: Summary of results for the several studies run with TDWB.

Study	All #	All1 #	Metrics (TP/FN/FP/TN) for All & All1 (same)	NoNSE #	Metrics (TP/FN/ FP/TN)	Overall Score for All1 & NoNSE (max115)
1	129	35	13/10/7/8	23	9/14/0/15	51 45
2	339	43	23/0/14/1	29	21/2/0/15	87 105
3	123	21	23/0/9/6	13	17/6/0/15	97 85
4	16	13	17/6/7/8	8	7/16/0/15	71 35
22	333	42	23/0/13/2	28	21/2/0/15	89 105
23	114	21	19/4/9/6	12	14/9/0/15	77 70
24	7	7	6/17/1/14	6	5/18/0/15	28 25

Table 3 shows the number of remaining patterns for this strategy & its associated metrics. By definition it will have a value of 0 for FP. The final column reports, for both the All1 & NoNSE strategies, evaluations of the resulting metrics, where the scoring function used is: number of TPs * 5 - number of FPs * 2; i.e., the same parameters are used as in the COVERAGE module to assess the "strength" of a pattern. (Recall these numbers are parameters provided on each run by the analyst.)

Figure 2: Screenshot of the Coverage module.

4.3 Analysis of the Studies

• The number of patterns reported for both Study-4 & Study-24 are both very low, this is because here we used a very small number of descriptor-value pairs; in fact only those at level-4 for the 3 descriptors (making a total of just 5 descriptor-values pairs).

⁷ The identified patterns reported here are able to predict the occurrence of PSEs, whereas in the Glasgow study (section 1), the patterns report associations between the identified descriptors and the PSE marker. (So in the

• On the other hand, we see that a larger number of descriptor-value pairs results in many more patterns being produced (Study-2 & Study-22).

• Not surprisingly, Study-22 produces a smaller number of patterns than Study-2, as the length of the temporal patterns returned here is more restricted.

• The relatively large number of patterns produced by Study-2 when, say compared with Study-3, results in a higher proportion of FPs being produced (both have 23 TPs and 14 & 9 FPs respectively).

• The previous points suggest that if the description space is too restrictive the coverage of the PSEs (i.e., TPs) is low, however if the description space is too large then all the PSEs will be covered but so will many of the NSEs. Studies 2, 3, 22 & 23 show a trade-off between these factors.

• What is of most significance here is that the coverage produced by Studies-2, 3, 22 & 23 are all better than that produced by Study-1, even though Study-1 used an additional descriptor, namely Fio2⁷.

5. DISCUSSION & FUTURE WORK

5.1 Discussion

The last point reports that the Apriori (AP) algorithm, which makes systematic searches through the data, produces larger number of patterns than the model formulated as a result of knowledge acquisition (KA) with the domain experts. However, many of the patterns produced by the AP algorithm might not be clinically acceptable i.e., they might describe dataset features which are considered by domain experts unlikely to precede a raised troponin value. This point needs to be investigated thoroughly.

A preliminary review of the patterns produced in Study-3 with a senior clinician, suggests that some of the patterns suggested are not acceptable but

case of association the order of the 2 entities is not significant, whereas in the case of prediction it is.) The issue of pattern type (associational or predictive) is also discussed at the beginning of Section 4.

some are. For instance the clinician accepted as a likely cause of MD several patterns which are less stringent than those inherent in the initial KA-derived model.

5.2 Future Work

Further Work will include the following tasks:

- Run systematic studies with domain experts to evaluate the clinical acceptability of patterns produced by TDWB; then capture their constraints as a TDWB filter which eliminates the unacceptable patterns
- Enhance TDWB's Coverage module so that it has a wider range of statistical functions to assess the strength of a proposed pattern; and introduce procedures to select the "best" set of patterns to fit an expert's criteria
- Link TDWB to appropriate ontologies to provide at least the domain terminology
- Use TDWB with further clinical datasets (e.g., the onset of diabetes, when to ventilate ICU patients); as well as ones from Ecology & Finance.

5.3 Acknowledgements

- TDWB was implemented by Sam Cauvin & Michael Gibson (University of Aberdeen) with financial support from the University of Aberdeen Development Trust
- Useful discussions on aspects of the design of TDWB with Dr Wamberto Vasconcelos (University of Aberdeen).

This work was an extension of the routine audit process in Glasgow Royal Infirmary's ICU; requirements for further Ethical Committee Approval have been waved.

6. REFERENCES

- Agrawal R. and Srikant R. 1994. Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94) (Santiago de Chile, Chile, September 12-15, 1994). Morgan-Kaufman, San Francisco, 487-499.
- Combi, C., Keravnou-Papailiou, E., Shahar, Y. 2010. Temporal Information Systems in Medicine. Springer, Heidelberg.
- Laxman S. and Sastry P.S. 2006. A survey of temporal data mining. SADHANA, Academy Proceedings in Engineering Sciences 31, 2, (April 2006), 173-198.
- Moss, L., Sleeman, D., Sim, M., Kinsella, J. Using Cardiovascular Derangements to Predict Raised Troponin Levels. 1st International Workshop on Capturing and Refining Knowledge in the Medical Domain (KMED 2012) <http://homepages.abdn.ac.uk/dcorsar/pages/kmed2012/papers.php>
- Sleeman, D., L Moss, M Sim & J Kinsella (2011). Predicting Adverse Events: Detecting Myocardial Damage in Intensive Care Unit (ICU) Patients, Proceedings of KCAP 2011 Conference, Publ: AM press, pp73-80.
- Sleeman, D., & D Blasco (2012). User Manual for the TDWB (Temporal Discovery Workbench) version 6.7.3, Technical report, Computing Science Department, University of Aberdeen.
- Sleeman, D., Aiken, A., Moss, L., Kinsella, J., Sim, M. 2009. A system to detect inconsistencies between a domain expert's different perspectives on (classification) tasks. In Advances in Machine Learning II, Studies in Computational Intelligence. Springer Berlin / Heidelberg, 293-314.