A ROBUST VALIDATION FRAMEWORK FOR TREND MINING: A STUDY IN DIABETIC RETINOPATHY

Vassiliki Somaraki[1, 2], Lee McCluskey[1]
[1] University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK
[2] St Paul's Eye Unit, Royal Liverpool University Hospital L7 8XP, UK

ABSTRACT

Data mining is concerned with the identification of hidden patterns in data. Trend mining is a branch of Data Mining that focuses in the process to indentify and analyze hidden trends in temporal data. A novel trend mining framework is described in this paper. The framework considers trends in terms of sequences of support values associate with frequent itemsets and uses a trend mining algorithm that produces prototypes trends. To validate the framework in the analysis of the generated trends a mechanism is also proposed. The framework is evaluated using longitudinal Diabetic Retinopathy screening data.

Keywords Validation Trend mining Longitudinal data Temporal databases

1 INTRODUCTION

Data mining is concerned with the identification of hidden patterns in data. The nature of both the patterns and data sets can take many forms. Pattern mining has wide applications in problems like association mining, classification or clustering (Hand D., (2001)). The mining of change points and trends across time stamped data sets is a challenging problem in data mining. One approach is to measure changes in patterns across the time stamps featured in the dataset. The patterns in this case can be defined as groups of attributes-values that frequently co-occur. Such groups of attribute-values are called items sets; if their frequency of occurrence is above some threshold they are called frequent item sets. Given a specific item set, its frequency of occurrence is known as its support .One mechanism for identifying frequent item sets is known as Association Rule Mining (ARM) (Agrawal et al (1994)). An Association Rule is a probabilistic relationship derived from identified frequent item sets. As such ARM can be thought of as a two stage method: (i) frequent items set identification, and (ii) AR generation. Association rule mining is a well researched method for discovering interesting relations between variables in large databases.

Trends across time stamped data sets can therefore be identified by observing the change in the support values of items sets across the data set. Trend mining is a branch of Data Mining that focuses in the process to indentify and analyze hidden trends in temporal data. The study described in this work is directed towards longitudinal patient data (longitudinal data is data collected using the same set of attributes at a series of points over time), more specifically at diabetic retinopathy screening data collected by St. Paul's Eye Unit, Royal Liverpool University. Diabetic Retinopathy (DR) is a common complication of diabetes -the most common cause of blindness in working age people in the UK. DR is a chronic disease affecting patients with Diabetes Mellitus and causes damage to the retina. (Kanski J., (2007)). Over 3,000,000 people suffer from diabetes and at least 750,000 people are registered blind or partially sight in the UK. The remainder is under the risk of blindness. Consequently it is important that Diabetic Retinopathy must be diagnosed at an early stage and accurately.

The research objective of the work is to investigate and identify a mechanism or mechanisms, whereby longitudinal data trends can be mined and the results presented in such a way that informed decisions can be made by policy makers, etc. Broadly this entails a number of issues:1)The mechanism for the pre-processing of the longitudinal data required to permit the desired trend mining 2)The nature of the trend mining mechanisms to be employed 3)The identification of the process to be used to present the results in a meaningful way. Longitudinal data thus provides a record of the "progress" of some set of features associated with the subjects. Medical longitudinal data, such as the DR data, typical plots the progress of some medical condition. Longitudinal data thus implicitly contains information concerning trends.

The work has resulted in a framework called "SOMA" which not only enables trend mining, but also supports the validation of discovered trends. This validation is based upon the selection of certain attributes for which there are known associations. Having known these associations as well as the patterns they change, it is possible to spot them through the prototypes that SOMA produces. Hence, the main purpose of this research is to develop a novel trend mining framework (SOMA) for extracting trends from longitudinal data while emphasizing validation of those trends.

This paper introduces the described method as a general framework for trend mining validation that can be applied generally to most trend procedures and types of data. The work also encompasses a number of issues: section 2 introduces trend mining, section 3 gives a description of the validation framework and section 4 includes the experimental evaluation of the application.

2 TREND MINING

Given the above definition of a trend we can adapt ARM techniques to identify all trends in a given data set. Piatetsky-Shapiro (1991) defined ARM as a method for the description, analysis and presentation of strong rules, called Association Rules (ARs), discovered in databases using different measures of interestingness. As noted above ARM is concerned with the discovery, in tabular databases, of rules that satisfy defined threshold requirements. Of these requirements, the most fundamental concerns the support (frequency) of the item sets used to make up the ARs: a rule is applicable only if the relationship it describe occurs sufficiently often in the data.

Dong et al (1999) defined Emerging Patterns (EPs) as an itemset whose support increases significantly from one dataset to another. EPs capture emerging trends in time-stamped datasets or capture differentiating characteristics between classes of data. However, the discovery of EPs in large datasets remains a challenge because of the large number of candidate patterns that may be generated. Dong et al (1999) defined the Growth Rate as follows: Let X be an item set and its support in a dataset D_1, $SUPP_{D1}(X)$ and in a dataset D_2 $SUPP_{D2}$. Given two datasets D_1 and D_2, the growth rate of an item set X from D_1 to D_2 is defined as:

$$GROWTH\ RATE(X) = \begin{cases} 0 & if\ SUPP_{D_1}(X) < \xi\ and\ SUPP_{D_2}(X) < \xi \\ \infty & if\ SUPP_{D_1}(X) = 0\ and\ SUPP_{D_2}(X) \neq 0 \\ \dfrac{SUPP_{D_2}(X)}{SUPP_{D_1}(X)}, & otherwise \end{cases} \quad (1)$$

Given Growth Rate threshold ρ> 1, an Emerging Pattern from D_1 to D_2 is an item set X where {Growth rate $D_1 \rightarrow D_2$ (X)} > ρ.

A Jumping Emerging Pattern (JEPs) is a special type of EP where the change in support between timestamps is infinite, we say that the support value “jumps”, for example zero frequency in D_1 and non-zero frequency in D_2 (Dong et al (2005))

For the work described here the TARM algorithm (Somaraki et al (2010 a)) was adopted; however effective ARM algorithms could equally well be used. TARM was selected because it uses a set enumeration tree structure, the T-tree, that facilitates fast “look-up”. The original ARM algorithm did not take the temporal dimension into consideration.

A trend can be described in a number of ways but the most obvious is as a time series plotting time against some value. From a longitudinal data trend mining perspective we are interested in identifying all “interesting” trends in the data. The definition of “interesting” is of course subjective but it can be worked on interestingness measures conducted in the field of Association Rule Mining (ARM), such as that of Hand et al(2001). The most commonly used interestingness framework used in ARM is the support-confidence framework, although it has its critics. ARM is concerned with the discovery of relationships between disjoint sets of attributes that feature in the input data. The relationships are defined in terms of Association Rules (ARs) of the form “if X occurs in a record then it is likely that Y also occurs” (where X and Y are disjoint subsets of some global set of attributes *A*). ARs are generated from identified frequent itemsets. A number of ARs may be generated from a single frequent itemset. A frequent itemset is a subset of *A* that occurs frequently in the input data. The frequency of a frequent itemset X is measured in terms of a support count, (s). A frequent itemset is deemed interesting if its support count is greater than some user specified threshold s. Consequently the number of generated frequent itemsets increases as the value of s decreases.

Trend Mining in Longitudinal Diabetic Retinopathy Data with respect to the work described here, and given the above, a trend t is defined as a time series describing the fluctuating support counts associated with an itemset over a sequence of n time stamps, thus t = $\{s_1, s_2, \ldots, s_n\}$. An approach to trend mining is to use the concept of user defined temporal prototypes to define the nature of the trends of interests. The trends are defined in terms of sequences of support values associated with identified frequent patterns. The prototypes are defined mathematically so that they can be mapped onto the temporal patterns. The trend mining framework SOMA, and Aretaeus, the associated trend mining algorithm have been developed. The proposed framework is able to detect different kinds of trends across longitudinal datasets.

The operation of the SOMA framework is the following: from the input of data, via the Aretaeus algorithm, to the final output. More precisely, the raw data first goes to the warehouse; and then to a New Data Pre-

processing Software where data cleansing, creation of data timestamps, selection of subsets for analysis and the application of logic rules take place. The data, after pre-processing, then goes to the data normalization stage, after which the frequent patterns are generated by applying an Apriori frequent pattern mining algorithm to every episode (defined by a unique time stamp) in the given data set. Then the frequent patterns and their frequency of occurrence are passed to Aretaeus algorithm to apply trend mining in order to produce different kind of prototype trends across the datasets based on the changes of the support (see figure 1). The Aretaeus algorithm uses mathematical identities (prototypes) to classify trends (see table 1).
As noted in the foregoing, given a low value for support threshold, a great number of trends are usually discovered to the extent that the number of identified trends hampers interpretation by decision makers. Potential solutions are to place constraints on the nature of the identified frequent itemsets, for example we may insist that we are only interested in frequent itemsets that feature certain attributes or combinations of attributes. The view taken in the work described here is that end users are interested in particular categories of trend. To this end distinct trend types have been identified and defined in the table below:

Type	Mathematical conditions
Increasing	$\frac{S_{i+1}}{S_i} > 1, \forall\, i \in \{1,2,...n-1\}$, GR>ρ
Decreasing	$\frac{S_{i+1}}{S_i} < 1, \forall\, i \in \{1,2,...,n-1\}$
Constant	$\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\, i \in \{1,2,...,n-1\}$, k : tolerance threshold
Fluctuating	$\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\, i \in \{1,2,...,n-1\}$ *and* $\frac{S_{j+1}}{S_j} > 1, \forall\, j \in \{1,2,...,n-1\}, j \neq i$ $\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\, i \in \{1,2,...,n-1\}$ *and* $\frac{S_{j+1}}{S_j} < 1, \forall\, j \in \{1,2,...,n-1\}, j \neq i$ $\frac{S_{i+1}}{S_i} > 1, \forall\, i \in \{1,2,...,n-1\}$ *and* $\frac{S_{j+1}}{S_j} < 1, \forall\, j \in \{1,2,...,n-1\}, j \neq i$ $\frac{S_{i+1}}{S_i} > 1, \forall\, i \in \{1,2,...,n-1\}$ *and* $\frac{S_{j+1}}{S_j} < 1, \forall\, j \in \{1,2,...,n-1\}, j \neq i$ *and* $\frac{S_{l+1}}{S_l} = 1 \pm k, \forall\, l \in \{1,2,...,n-1\}\ l \neq j, l \neq i$
Jumping	*for* $m < n$: $S_i = 0, \forall\, i \in \{1,2,...,m\}$ *and* $S_i > 0\ \forall\, i \in \{m+1,n\}$
Disappearing	*for* $m < n$: $S_i > 0, \forall\, i \in \{1,2,...,m\}$ *and* $S_i = 0\ \forall\, i \in \{m+1,n\}$

Table 1 : Mathematical identities for the trend mining application

3 TREND MINING VALIDATION FRAMEWORK

The above mechanisms have been implemented into a trend mining framework for application to longitudinal data. The framework takes as input a data set D = $\{d_1,d_2,\ldots,d_n\}$ and a support thresholds and produces a set of trends categorized according to the above. Each dataset d comprises a sequence on m records. Each record has associated with it a set of attributes which is some sub-set of the global set of attributes *A*. The framework incorporates an Apriori algorithm, as described above; and the trend mining mechanism for conducting the desired categorization. Further details concerning an early implementation of the trend mining framework described here can be found in (Somaraki et al (2010 a)) and (Somaraki et al (2010 b)). In this section we describe an extended SOMA framework which incorporates the validation stages outlined above. This forms a generic solution for validation so that the framework can be transferred to other similar studies where there is:

- data which is temporal
- access to a model (e.g. a set of rules) capturing the constraints of the scientific area
- a set of expected or known associations.

We aim the testing of SOMA towards finding associations among the data that are already known by the applications experts.
Among the attributes that have been selected, there should be at least one that has the role of the “key attribute”. In this research a “key- attribute” could be an attribute that characterizes the status of Diabetic Retinopathy for each patient. The other attributes play the role of variables, and are called “variable-attributes”.
Therefore their values affect the value of the “key- attribute”. Suppose having a mathematical multivariate function F(x, y, z... w), where the value of the function F (diabetic retinopathy) is determined from the values of the variables x, y, z... w (for example: age, diabetes duration, diabetes treatment etc).
Each trend that is generated based on the mathematical identities contains the following pieces of information:

- the name of the attributes
- the value of the attribute
- the frequency (support value) at each time stamp

The output from the trend mining algorithm for each time stamp (episode) consists of many itemsets. Their number is determined by the support threshold, which is the minimum frequency that every pattern must reach and overcome in order to be recognized as a “frequent pattern”, and be passed to the Aretaeus software which performs the trend mining. Therefore it is clear that not all trends will contain all the information that is needed, for example there will be trends that they will contain only “variable-attributes”. Hence in the validation process we will use only the trends that definitely contain a “key attribute” and at least one “variable- attribute”. The number of those trends is strongly related with the support threshold, that there must be enough volume of information regarding these attributes. The information from the medical experts is a set of criteria where each contains a “key-attribute”, some “variables-attributes” and characterizes some known association between them.
An example of known association as given from experts to validate has the following form:
IF {RERET=10} THEN {calculated diabetes duration = short} & {Present treatment = 1}.In this example {RERET=10} is the key- attribute while {calculated diabetes duration = short} and {Present treatment = 1} are the variable-attributes.
The trend mining algorithm has been extended to incorporate an automatic check of the known association over the whole trend mining results.

4 EXPERIMENTAL RESULTS

To evaluate the trend mining framework it is very important to check the validity upon known associations. In large databases with many attributes the combinations between the attributes could be numerous. Therefore, there is an issue whether the associations that produced are valid and how to measure their validity. An approach for this is to examine whether the framework can reproduce known association. If it is successful it can be considered that any other association is valid and true.
In order to address this problem the SOMA framework it is tested against known associations. These associations are come from the experts and they are combinations of the values of some attributes. The goal is to confirm:
i) whether SOMA can give as output these association and
ii) that the confidence of the association is above certain limit.
This section presents an evaluation of the trend-mining framework introduced in the foregoing. The evaluation was directed at an analysis of: (i) the number of trends that might be discovered and (ii) the nature of the classification trend. Changing the support threshold we observe the changes of the number of trends for each different category.
The Diabetic Retinopathy database used in the evaluation was extracted from a number of databases maintained by The RLUH as part of its diabetes screening program. Overall RLUH has recorded details of some 20,000 patients spanning an eighteen year period. Patients with diabetes are usually screened once a year. Patients enter and leave the screening program at different times; the average time that a patient spends within the screening is currently six years. Thus, for the evaluation, only those patients that had taken part in the program for at least four years were selected. Where patients had been in the program for more than 18 years, data from the first four consultations was selected. The dataset of this experiment comprises six time stamps with 1430 records per timestamp. 7 data attributes were used for the evaluation including the level of diabetic retinopathy for both eyes, the kind of treatment, the duration of the disease

from the date it was firstly diagnosed, age of the patient, age of the patient when he or she was first diagnosed,the type of diabetes; which, after normalization and discretization, resulted in 215 attributes-values. It is worth noting that the data required significant “cleaning” to remove noise and address various anomalies; however, the nature of this data cleaning is beyond the scope of this paper. Table 2 presents how the change in the support threshold affects the number of trends.

Support Threshold %	Total trends	Increasing trends	Decreasing trends	Fluctuating trends	Jumping trends	Disappearing trends
0.5	7602	14	25	930	1376	1827
1	2532	12	12	638	559	714
2.5	874	1	2	134	193	235
5	266	0	3	11	59	74
10	108	0	6	3	25	25

Table 2 Evaluation of SOMA framework against different support threshold values

As regards validation process the following rule was tested:
Present_Treatment=1 calculated_diabetes_duration=1 --> RE_RET=10
The validation framework gave the following results: S={402 293 188 80 36 0} where S is the vector of the support values (number of patients) for every time stamp.
The kind of trend is *disappearing* and the Validation rule has maximum confidence: 96.8675 % and minimum confidence: 0.0000 %.Here confidence can be interpreted as the probability of finding the key attributes (RHS) of the rule in the records under the condition that these records also contains the variable attributes (LHS). Because the trend is disappearing the min confidence is 0. At the last time stamp it can be seen that the support value is 0. That has the explanation that the number of records is less that the support threshold. Generally, for any rule given, in order to be valid must have confidence equal or greater than the minimum confidence threshold.

5 CONCLUSIONS

A mechanism for identifying, analyzing and validating trends has been described. The mechanism incorporates the extension of SOMA framework including a robust external validation algorithm. The mechanism has been realized in the form of a trend mining framework which has been applied to Diabetic Retinopathy screening data. As part of future work there will be an investigation of a generic approach at each of the beginning/end of the constituent processes. A language for a set of declarative validation rules will be set up, and a systematic process of validation of each set of input data will be created.

REFERENCES

AGRAWAL, P., SRIKANT, R. (1994). *Algorithms for mining Association Rules Proceedings of the 20th International Conference on Very Large Databases* pp. 487-499.

DONG, G., LI, J. (1999). *Efficient Mining of Emerging Patterns :Discovering Trends and Differences.* 5th International Conference in Knowledge Discovery and Data Mining pp. 1-11.

DONG, G., LI, J. (2005*). Mining border descriptions of emerging patterns from dataset pair. Knowledge and Information Systems*. 8(2) 178 - 202.

FAN, H., RAMAMOHANARAO, K. (2003). *A Bayesian Approach to Use Emerging Patterns for classification*. Proceedings of the 14th Australasian database conference pp. 39-48.

FAN, H. (2004*). Efficient mining of interesting Emerging Patterns and Their Effective use in classification*. PhD Thesis.The Department of Computer Science and Software Engineering.University of Melbourne, Melbourne.

HAND, D. (2001*). Principles of Data MIning* MIT Press pg 204-205.

KANSKI, J.(2007). *Clinical ophthalmology: A systematic Approach*. 6th edition, Elsevier.

PIATETSKY-SHAPIRO, G. (1991) *Discovery analysis and presentation of strong rules*.In Frawley, G. P.-S. W. J. (ed), Knowledge Discovery in Databases.

SOMARAKI V., BROADBENT D., COENEN F., HARDING P.S.(2010 a). *Finding Temporal Patterns in Noisy Longitudinal Data: A Study in Diabetic Retinopathy*. ICDM 2010: 418-431.

SOMARAKI V., HARDING P S., BROADBENT D., COENEN F. (2010 b). *SOMA: A Proposed Framework for Trend Mining in Large UK Diabetic Retinopathy Temporal Databases*. SGAI Conf. 2010: 285-290.

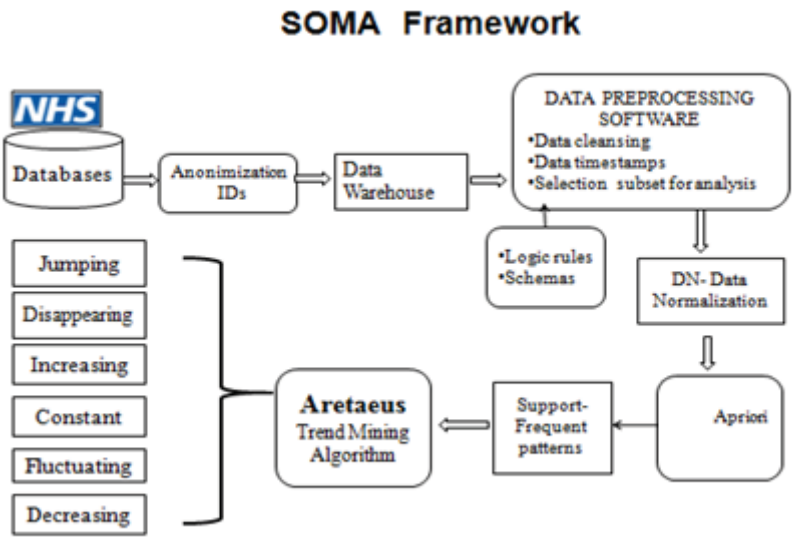

Figure 1: Representation of SOMA Framework