Mining Medical Data for Predictive and Sequential patterns: PKDD 2001

Susan Jensen, SPSS UK Ltd

St Andrew's House
West Street
Woking, Surrey GU21 1EB
+44 1483 719 296
sjensen@spss.com

Abstract. Data relating to patient information and medical exams connected with thrombosis attacks were analysed using SPSS Clementine data mining workbench. The ability to predict the onset and successful diagnosis of thrombosis is key to the intervention of the disease, and sequential patterns of symptoms and lab exams may indicate a trending from a pre-thrombosis to active thrombosis condition. In this report, predictive modelling, association rules and sequence detection were used to investigate these patterns.

1 Introduction

Thrombosis is a potential complication of collagen diseases. In pursuit of understanding the mechanisms responsible for collagen diseases and the onset of thrombosis, analyses have been undertaken to quantify a profile of the lab exam results that indicate thrombosis presence, and the sequence of a) symptoms b) exam results and c) types of collagen diseases that can predict or be associated with the onset of thrombosis. The PKDD 2001 challenge presented several relational tables with which to explore the patterns involved in the onset of thrombosis.

Effectively mining large data sets is best facilitated by a systematic process. Using the cross-industry standard process for data mining (CRISP-DM, www.crisp-dm.org), supported by Clementine, recommended steps are clearly outlined, from understanding the problem through to deploying results back into the business or organisation. The remainder of this report will be structured generally in the CRISP-DM format.

2 Business understanding

The nature of the Challenge dictates that, unusually, domain knowledge isn't necessarily a prerequisite. Lack of in-depth knowledge on the part of the author about the medical condition and about the relationship between lab exam results and the

condition, preclude stating whether the results have significant medical value, but may point to strong directions for further research or ways in which algorithms may be employed to extract patterns from medical data.

As such, the general questions addressed here will be what factors may be predictive of the onset of thrombosis, and whether there are sequential or associated indicators of lab results or other collagen diseases that may point to or predict the onset of thrombosis.

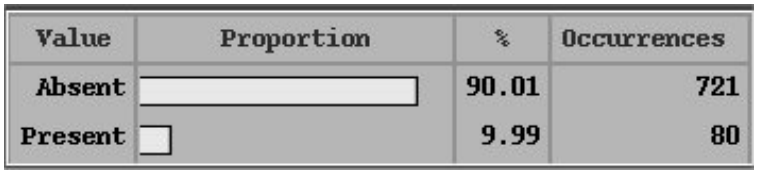
3 Data understanding

The data were provided by PKDD, and the flat files were put into an Access database for ease of use.

Tables used to extract information:

- Changes in lab results over time: Lab_exam
- Patient demographics and primary ID key: Patient_info
- Thrombosis diagnosis: Spc_exam
- Symptoms leading to Thrombosis: Other_symptoms
- Other collagen diseases: Diagnostic

The field containing diagnosis of Thrombosis was grouped into a Thrombosis Present/Absent field for visualisation and prediction purposes, and a separate category indicating those for whom no thrombosis diagnosis was made. Initial examination indicates that of all patients in the database for suspected or confirmed collagen diseases, about 10% of patients were diagnosed as having one of three forms of thrombosis (most severe, severe, mild) present (Fig. 1).



Value	Proportion	%	Occurrences
Absent		90.01	721
Present		9.99	80

Fig. 1. Proportion of thrombosis present in patients assigned a diagnosis

4 Data Preparation and Modelling

As the relationship between these two steps in the data mining process is highly iterative, I include both in the same section for brevity of this document.

Much data cleaning and reformatting was done to make analysis possible and more interpretable, including reformatting of dates to a consistent type, extracting numeric information from numeric fields that contain a few symbolics (ie <5000 as an entry), etc. One fortunate aspect of Clementine is that all the manipulations are clearly visible in the stream, contained in the series of nodes leading from the data source node(s) to the output nodes, which may be a table, chart or series of models. In Clementine, as a data mining workbench, a visual record of the data mining process is self-maintaining.

The topics addressed are contained in the following sections.

4.1 Understanding factors that lead to onset of thrombosis – profiles of lab exams

Numeric fields were banded into low/medium/high individually, using interactive histograms. The bands were created to have approximately the same number of individuals in each. Where numerics fields were given that contained symbolic exceptions (such as <31.2), the '< ' was stripped and the (value –1) was arbitrarily used in its place. This has inherent problems but was used as a reasonable general guide. Where no result was given, an 'empty' was assigned to the field.

A backpropagation neural network was run on all possible variables of the lab exams, to see what results were most indicative of thrombosis presence/absence. Where numeric variables contained symbolic exceptions, the numbers were extracted so that all possible variables would be treated as numerics. The symbolic variables were left as originally presented for use in the model.

Table 1. Misclassification rate of neural network predicting presence/absence of thrombosis based upon lab exam results

Results for output field ThrombosisFlag

Comparing $N-ThrombosisFlag with ThrombosisFlag

Correct	: 17916	(88.21%)
Wrong	: 2395	(11.79%)
Total	: 20311	

Coincidence Matrix

	$N-ThrombosisFlag	
	Absent	Present
Absent	13490	1412
Present	983	4426

Neural network sensitivity analysis output lists all variables included in the model in a relative order of importance, based upon an internally derived algorithm. The high scoring variables contribute most, lower scoring are of decreasing value to the model.

Table 2. Sensitivity analysis of neural network, indicating relative importance of variables included in model creation.

Neural Network "ThrombosisFlag" architecture	
Input Layer	: 96 neurons
Hidden Layer #1	: 11 neurons
Output Layer	: 3 neurons
Predicted Accuracy	: 85.82%
Relative Importance of Inputs	
LAC	0.31256
ANA	0.28393
U-PRO	0.23903
CENTROMEA	0.21101
SSA	0.20978
SSB	0.19812
RNP	0.19050
SM	0.18970
SC170	0.16079
ACL IGG	0.15170
KCT	0.12847
RVVT	0.12842
AT3num	0.11620
ACL IGA	0.10687
ACL IGM	0.09802

The factors with a Relative Importance of 0.150 (Table 2) and above (LAC, ANA, U-Pro, Centromea, SSA, SSB, RNP, SM, SC170) were used in the lab exam sequence analysis (next pages).

Rule induction (C5.0) was used to profile presence/absence of thrombosis among the banded numeric variables, looking for rules that had at least 300 individuals in each decision tree branch, and with a strong pruning severity imposed on the tree. Rules with high classification confidence were uncovered (Table 3)

The C5.0 ruleset (Table 3) lists profiles of the target variable in order of confidence (percent of that profile that reaches the conclusion, in this example of ‘Present’) and coverage (the number of cases that contributed to creating that rule).

For example, 2000 individuals are described by the rule that where the ATTP result is high, there is an 87% chance that they will develop thrombosis, assuming that the training population is representative.

The decision tree generated with the above ruleset indicates that of the banded numeric variables, the most important contributors to predicting onset of thrombosis were ATTP, IGG, WBC. Figure 2 presents a lower part of the text decision tree,

Table 3. Rule set describing characteristic lab exam results of individuals diagnosed with thrombosis. Numbers in brackets indicate (number of individual patients making up that rule, percent confidence with which outcome can be applied to that set of characteristics)

```
Rules for Present:
   Rule #1 for Present:
      if  APPTband == high
      then -> Present (2000, 0.867)

   Rule #2 for Present:
      if GLUband == empty
      and WBCband == high
      and PLTband == low
      and RFband == empty
      and IGGband == empty
      then -> Present (1330, 0.77)

   Rule #3 for Present:
      if  PLTband == medium
      and APPTband == empty
      and IGGband == low
      and GPTband == high
      then -> Present (372, 0.693)

   Rule #4 for Present:
      if  APPTband == empty
      and IGGband == low
      and GPTband == high
      then -> Present (821, 0.691)

   Rule #5 for Present:
      if  GLUband == empty
      and PLTband == low
      and RFband == empty
      and IGGband == empty
      and GPTband == medium
      then -> Present (1314, 0.683)
…
```

where the most important factors are further to the left and less important factors are nested within, until all the conditions leading to a conclusion (thrombosis Present/Absent) are reached (Fig. 2). As with the rule set, the numbers within the brackets indicate the number of cases that make up the branch/leaf, and the confidence with which that branch or leaf predicts the scored outcome.

The sensitivity analysis of the neural network and the order most significant factors in the decision tree indicate those factors in the lab results that might be worth looking at in more detail. The question addressed next was: among those lab results that were most predictive of the above models, are there any patterns in them over the temporal sequence of lab exams that could be used as indicators of onset of thrombosis? Or: what are typical patterns in collagen disease patients over time?

4.2 Analysis of sequence of lab exam results

Sequential analysis was performed using CaPri. Rules were set to find sequences of length 2-5 using the banded variables of APPT, IGG and WBC. Rules were set to find sequences where at least 50, 20 or 30% of the population (WBC, APPT, IGG respectively) were represented in each sequential rule, and where the confidence of

each rule indicating that the final step in the sequence would occur set at 80, 80 and 70% respectively. In short, very stringent stopping rules were applied to show only the most popular sequences.

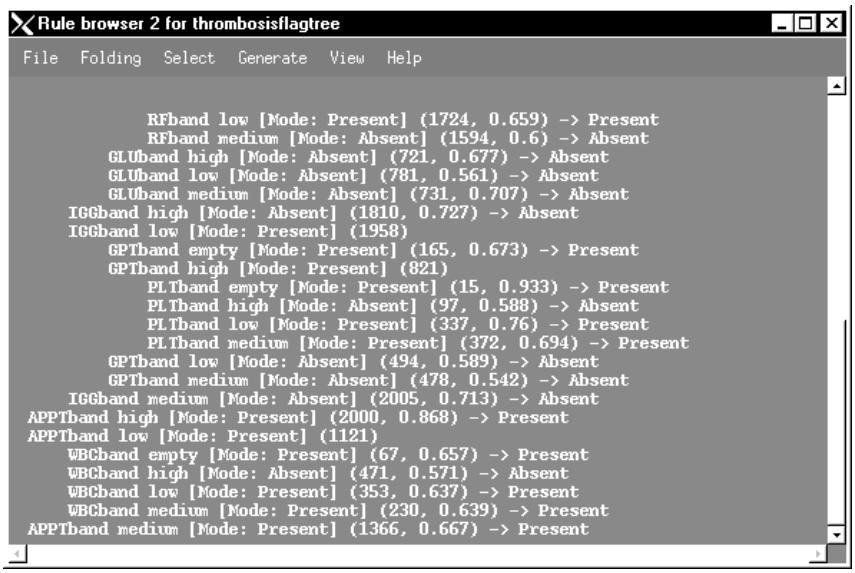


Figure 2. C5.0 rule induction decision tree, text version. Variables to the left are of higher importance in distinguishing between thrombosis being Present or Absent. Tree is output in multiple colours, presented here in greyscale for publication.

In order of appearance for each rule number, in the brackets of CaPri output (Figs 3, 4, 5), the figures state: Sequence length, occurrence, support, confidence. For example, look at the WBC set of rules (Fig. 3). Number 6 says that of a three-exam sequence length, those three items occurred 251 times in the table, and covers 60.34% of the database. The confidence states that of all those individuals who had the first two exam results (medium-medium, in 292 cases, seen in rule 2), 85% went on to the third exam result in the sequence (low).

Looking only at the very strongly represented sequences (Figs 3, 4, 5), none were more likely than others to fluctuate wildly over the time period. IGG (Fig. 5) appears to remain higher in presence of thrombosis than the other two exam results. A look through further lab exam results that might have a physiological basis for fluctuation in a thrombosis patient might show warning patterns, or set a baseline for the norm.

1 (2, 285, 68.51%, 84.32%) medium, low
2 (2, 292, 70.19%, 86.39%) medium, medium
3 (2, 333, 80.05%, 91.99%) low, low
4 (2, 250, 60.10%, 88.34%) high, high
5 (3, 266, 63.94%, 93.33%) medium, low, low
6 (3, 251, 60.34%, 85.96%) medium, medium, low
7 (3, 269, 64.66%, 92,12%) medium, medium, medium
8 (3, 308, 74.04%, 92.49%) low, low, low
9 (3, 266, 63.94%, 92.68%) low, medium, low
10 (3, 256, 61.54%, 89.20%) low, medium, medium
11 (4, 294, 70.67%, 95.45%) low, low, low, low
12 (5, 280, 67.31%, 95.24%) low, low, low, low, low

Figure 3. CaPri (sequence detection) rule set for the series of lab exams of WBC levels over time, where the numeric information has been banded into low, medium and high levels. See text for explanation of rule contents.

1 (3, 48, 22.97%, 96.00%) high , high , high
2 (4, 45, 21.53%, 93.75%) high , high , high , high
3 (4, 43, 20.57%, 84.31%) medium , low , low , low
4 (4, 43, 20.57%, 86.00%) medium , medium , low , medium
5 (4, 55, 26.32%, 80.88%) medium , medium , medium , medium
6 (4, 43, 20.57%, 87.76%) low , medium , medium , medium
7 (4, 62, 29.67%, 88.57%) low , low , low , low
8 (5, 45, 21.53%, 81.82%) medium , medium , medium , medium , medium
9 (5, 53, 25.36%, 85.48%) low , low , low , low , low

Figure 4. CaPri (sequence detection) rule set for the series of lab exams of WBC levels over time, where the numeric information has been banded into low, medium and high levels. See text for explanation of rule contents.

1 (2, 132, 34.46%, 77.19%) low , medium
2 (2, 131, 34.20%, 76.61%) low , low
3 (2, 205, 53.52%, 82.00%) high , high
4 (2, 224, 58.49%, 80.87%) medium , medium
5 (3, 124, 32.38%, 80.00%) high , medium , medium
6 (3, 170, 44.39%, 82.93%) high , high , high
7 (3, 117, 30.55%, 81.82%) medium , high , high
8 (3, 179, 46.74%, 79.91%) medium , medium , medium
9 (4, 154, 40.21%, 90.59%) high , high , high , high
10 (4, 153, 39.95%, 85.47%) medium , medium , medium , medium
11 (5, 128, 33.42%, 83.12%) high , high , high , high , high
12 (5, 130, 33.94%, 84.97%) medium , medium , medium , medium , medium

Figure 5. CaPri (sequence detection) rule set for the series of lab exams of IGG levels over time, where the numeric information has been banded into low, medium and high levels. See text for explanation of rule contents.

A further step would be to see whether a particular sequence would lead to a thrombosis-present result. This was not done due to time limitations.

4.3 Association of symptom in patients with thrombosis present

Tracing symptoms over a patient history may indicate symptoms that tend to go together or regularly follow each other in thrombosis patients, and so analysis was performed using the association rules and CaPri algorithms. The first tells what tend to co-occur, the second uses timing information to indicate what tends to follow another symptom. Date manipulations were required to ensure accurate sorting by ID and Date for this step, and the fields used were ID and Symptom. Only 198 records were available for this analysis, so rule conditions and conclusions contain small numbers.

All symptoms were allowed to be both inputs and outputs, so there is repetition in the association, since rules generated disregard sequential occurrence. Output from Apriori (association rules) gives conditions (to the right of the arrow) leading to a conclusion (to the left of the arrow) (see Fig. 6). In the brackets, the figures indicate the number of cases making up that association (and the same cases may be repeated in different associations, if they are just flipped condition-conclusion), the percent of the population that made up that association, and the confidence with which the patient who has the condition(s) will go on to the conclusion.

```
Abortion < = Edema (5:2.5%, 0.5)
DVT< = CVA(APO) (5:2.5%, 0.5)
DVT< = Pulmonary Thrombosis (5:2.5%, 0.5)
DVT< = Edema (5:2.5%, 1.0)
Epilepsy < = CVA(APO) (5:2.5%, 0.5)
Epilepsy < = Edema  (5:2.5%, 0.5)
Thrombocytopenia < = CVA(APO) (5:2.5%, 0.5)
Thrombocytopenia < = Thrombocytopenia (5:2.5%, 0.5)
Thrombocytopenia < = Edema  (5:2.5%, 0.5)
Brain Infarction < = CVA(APO) (5:2.5%, 0.5)
Edema < = CVA(APO) (5:2.5%, 0.5)
CVA(APO) < = Edema  (5:2.5%, 0.5)
Abortion < = DVT & Edema (5:2.5%, 0.5)
Epilepsy < = DVT & Edema (5:2.5%, 0.5)
Thrombocytopenia < = DVT & Edema  (5:2.5%, 0.5)
    CVA(APO) < = DVT & Edema (5:2.5%, 0.5)
```

Fig. 6. Apriori association rules. Right side of arrow lists the condition(s) leading to the conclusion on the left side of the arrow. See explanation in text for numbers in brackets.

The highest confidence association comes from the five individuals from whom the condition of edema was always followed by a conclusion of DVT (Fig. 6). The number of cases in each association are a bit suspect, and require further investigation.

The Apriori result is reflected in the CaPri output (Fig 7). Symptoms generally followed each other, but in the case of edema, DVT would follow. Again, the small numbers make generalisation difficult but might present a warning to a physician.

```
1 (2, 2, 2.63%, 100.00%) PULMONARY THROMBOSIS , PULMONARY THROMBOSIS
2 (2, 2, 2.63%, 100.00%) LEG ULCER , LEG ULCER
3 (2, 2, 2.63%, 100.00%) MULTIPLE THROMBOSIS , MULTIPLE THROMBOSIS
4 (2, 2, 2.63%, 100.00%) EDEMA , DVT
5 (2, 5, 6.58%, 100.00%) EPILEPSY , EPILEPSY
6 (2, 4, 5.26%, 80.00%) THROMBOPHLEBITIS , THROMBOPHLEBITIS
7 (2, 5, 6.58%, 83.33%) ABORTION , ABORTION
8 (2, 5, 6.58%, 83.33%) CVA , CVA
9 (2, 16, 21.05%, 94.12%) CNS LUPUS , CNS LUPUS
10 (2, 5, 6.58%, 100.00%) AMI , AMI
11 (2, 4, 5.26%, 80.00%) PH , PH
12 (2, 6, 7.89%, 75.00%) DVT , DVT
13 (3, 3, 3.95%, 60.00%) ABORTION , ABORTION , ABORTION
14 (3, 2, 2.63%, 12.50%) CNS LUPUS , CNS LUPUS , CNS LUPUS
```

Fig. 7. CaPri (sequence detection) results of symptoms that tend to follow each other over time. Rule 4 echos results seen by Apriori association rules, but CaPri shows the order in which they appear.

4.4 Association rules, of collagen disease diagnosis and presence/absence of thrombosis

To examine the relationship of collagen disease diagnoses and the potential for association with thrombosis, the Apriori algorithm was used. Generally association rules are looking for anything as input/output (condition/conclusion), but in this case the presence of thrombosis was set as the conclusion and diagnoses the inputs, to direct the rule generation toward those diagnoses that lead to a conclusion of thrombosis. This is similar to the rule-induction rule sets, but a way of looking at the stability of the rules using another algorithm

```
ThrombosisFlag_Present < = diagnosis_APS (42: 9.7%, 0.619)
ThrombosisFlag_Present < = diagnosis_SLE & diagnosis_APS (19: 4.4%, 0.526)
ThrombosisFlag_Present < = diagnosis_SJS & diagnosis_APS (7: 1.6%, 0.571)
   ThrombosisFlag_Present < = diagnosis_SLE & diagnosis_SJS & diagnosis_APS (42:
9.7%, 0.619)
```

Figure 8. Apriori association rules indicating relationship of collagen disease diagnoses with presence of thrombosis. See text for explanation of numbers in brackets.

Closer examination of the different types of thrombosis that could be diagnosed showed that the numbers were too scarce to associate with anything but type 1 thrombosis, the most severe form.

Thrombosis_1 < = diagnosis_APS (42: 9.7%, 0.595)
Thrombosis_1 < = diagnosis_SLE & diagnosis_APS (19: 4.4%, 0.526)
Thrombosis_1 < = diagnosis_SJS & diagnosis_APS (7: 1.6%, 0.571)
Thrombosis_1 < = diagnosis_SLE & diagnosis_SJS & diagnosis_APS (5: 1.2%, 0.4)

Fig. 9. Apriori association rules with types of thrombosis present. Only type 1 thrombosis had sufficient numbers for associations to be detected. See text for explanation of numbers in brackets

5 Model Evaluation and Deployment

The validity of the models is difficult to judge due to lack of domain knowledge on the part of the author, but common sense and a test data set would be used at this point to evaluate whether the results are worth deploying into a medical organisation. In the context of this report, the conclusions tentatively reached include:

- LAC, ANA, U-Pro, Centromea, SSA, SSB, RNP, SM, SC170 were strong contributors to predicting the presence or absence of thrombosis, using a backpropagation neural network. On the training data, there was an overall predictive accuracy of 88%, with an approximately 82% accuracy of predicting presence of thrombosis based upon lab exams.
- Using the interactive graphs to manipulate the numeric lab exam results into banded information for ease of interpretability, profiles of patients with thrombosis were derived. For a patient with high levels of APPT, there was an 86% chance of thrombosis onset.
- Of the three lab exam results investigated, a sequential analysis did not indicate any patterns of regular rise or fall over the series of exams, nor did any of them show wildly fluctuating patterns. One tended to remain high more than the others over time.
- Association of symptoms indicated several patterns on a small data set, the strongest of which (in terms of confidence in the rule) indicated that edema was associated with DVT.
- Sequential analysis over a series of symptoms reflect the association rule above. Where most symptoms tended to remain static, the symptoms of edema would later be revised to that of DVT.
- Analysis of collagen disease diagnosis, and presence of thrombosis, indicates that over half the incidences of APS are associated with presence of thrombosis (of any severity). There are lower but perhaps warning patterns of other collagen diseases that become associated with thrombosis.
- Breaking down the different types of thrombosis, the above pattern was found to be dictated by the most severe form (type 1), largely due to the overwhelming numbers of type 1 relative to the others in the data set.