pKDD 2001 Discovery Challenge - Medical Domain

Ivan Coursac Nicolas Duteil Noël Lucas

co Pertinence Data Intelligence
4 rue Louise Michel
92300 Levallois-Perret
France

Abstract

This document presents a first analysis of clinical data collected at Chiba Hospital, concerning collagen diseases and thrombosis cases. The main goals of the study reside in the prediction of each patients'thrombosis class, given some available medical records, as well as the identification of the most relevant measurements to predict a given thrombosis class. The focus of the paper is to show that findings of medical interest can be discovered using data mining. Attention is drawn first on the data preparation phase, because of the particularity of the temporal aspects of the data. Three examples of discovered knowledge are then presented and discussed.

1 Introduction

Collagen diseases are auto-immune diseases, thrombosis being one of their most important and severe complications, one of the major causes of death. It is important to detect and predict the possibilities of its occurence. A major challenge for Data mining is to discover unknown regularities, hidden in the wealth of clinical data such as the data collected at Chiba Hospital. Discovered regularities will hopefully allow physicians (and patients) to:
-better prevent acute crises;
-request invasive or expensive exams when needed
-bring some understanding of the causes and side-effects entangled in the disease phenomenon.
In the following, we briefly present the data available in section 2, then the goals we set ourselves in section 3, our mining techniques to prepare the data for our analysis in section 4, and three examples of the discovered knowledge in section 5. The medical interest of such research in general, and of our research in particular, is discussed in the last section.

2 Presentation of data tables

The following table summarizes what kind of data we have for each type of patient:

	Lab.exams	Spc.exams	Thrombosis class
Type I	available		0
Type II	available	available	0,1 or 2
Type III		available	0,1 or 2

Table 1: Availibility of data by type of patient.

Lab.exams: standard lab tests, routine checks, non-specific.
Spc.exams: tests done to detect collagen diseases, and possibly thrombosis. These are costly.
Thrombosis classes: 0 is for healthy patients, 1 is for patients who have thrombosis, 2 for patients who develop thrombosis in the future.

3 Objectives, goals of experiments

Two goals will be considered in the following:

1. The first one is concerned with the prediction of the three selected classes of thrombosis , using the different measurements available (lab.exams and spc.exams)
2. The second one is concerned with identifying the most relevant measurement (specific exam, or lab.exam) in order to predict a given thrombosis class.

These goals can be best understood using a non medical analogy. The limits of our analogy will also serve the purpose of identifying more clearly the difficulties of our goals. Let us compare a patient to a bridge. Spc.exams give us an estimate of the state of fragility/solidity of the bridge: we can view them as the pillars of the bridge. Positive Spc.exams correspond to a risk of bridge collapse (medically speaking, this part of the analogy holds perfectly). A thrombosis attack corresponds to a bridge collapse. We could then compare the Lab.exams to all types of vehicles using the bridge.

We can then assume:

- IF a given type of vehicle, or combination of vehicles, pull out onto the bridge (i.e a given lab.exam takes a particular value, or a combination of lab.exams vary in a certain way)
- AND the bridge is weakened in some way (variations of Spc.exams)

- THEN the bridge collapses (or has collapsed: thrombosis class 1 or 2)

Following this analogy, our goals can be reformulated as:

1. Given the traffic and the general condition of the pillars, can we predict the collapse of the bridge?
2. Given the traffic, and knowing that the bridge has collapsed (or will collapse), can we identify which pillar gave in (or which pillar to watch in particular)?

4 Experimental setting

4.1 pre-processing, cleaning up, preparation

As emphasized by Fayyad [1], data preparation is one of the most important data mining task. The first stage in our data preparation phase consisted in cleaning the different data tables of all inconsistencies.

Some recoding was done for the date informations. The patient's birthdate was converted into an integer (number of days). Another date was taken as reference, the "age of the disease", intended as the first day where a lab exam was performed. Each exam is thereby attached an integer attribute, the number of days elapsed since the first exam.

The problem domain at hand present two particular difficulties. one is that different sets of patients are decribed using different attributes (Table 1). A co-training-like approach (Blum and Mitchell [2]) is on-going to address this problem.

A second difficulty is that there is a varying number of lab.exams for all patients. In order to apply attribute-value learners, the labs series was propositionalized by "unfolding" the patient case as follows:

Assume that patient John underwent a series of lab exams

$$y(time = t1), ..., y(time = t18)$$

This information is accounted for by considering instead patients John_1,...John_16, representing John at different time steps. For instance, John_1's description involves the lab exams at time t1, and the lab exams at time t2; the associated conclusion is John's health state at time t3.

After our "unfolding" phase, the data for one patient is organized in the following manner:

John ID	t	Lab.exams(t)	t+1	Lab.exams(t+1)	t+2	Lab.exams(t+2)	Spc.exams	Thromb.class(t+2)
John ID	t+1	Lab.exams(t+1)	t+2	Lab.exams(t+2)	t+3	Lab.exams(t+3)	Spc.exams	Thromb.class(t+3)

5 Examples of discovered knowledge

All experiments reported in the following were done using C5 (Quinlan [3]).

The experimental results were obtained using the 5 times 2 cross validation technique, after the validation procedure recommended by T.Dietterich[4].

5.1 Standard classification on Type II patients

Input: all type II patients, oganised in the "unfolded" table as shown earlier.
Result: prediction of the thrombosis class, with an overall 99.28% correct prediction rate. Following is the obtained confusion matrix:

	$class0$	$class1$	$class2$
$class0$	8144	5	7
$class1$	26	58	11
$class2$	17	3	1289

5.2 Informative nature of spc.exams

The informative nature of spc.exams was tested by comparing the results with spc.exams (reported in the previous subsection) and without spc.exams.
So, taking the "unfolded" table of type II patients and truncating it to exclude the Spc.exams, then joining the result table with the "unfolded" table of type I patients, and using that as a training set, we tried to predict the thrombosis class of all type II patients (validation set).
Although the overall result in terms of correct predictions is poor (only 86%), we were pleased with the results obtained on class1 patients, as shown in the following confusion matrix:

	$class0$	$class1$	$class2$
$class0$	16179	0	0
$class1$	185	15	1
$class2$	2570	0	157

Each time the algorithm classified a patient in class1, that patient truly had thrombosis. So what we have here is a specific test for class1, a specific test Sp being defined formally as follows:

$$\text{Sp} = \frac{TN}{TN+FP}$$

where TN stands for 'true negative', and FP stands for 'false positive'.

In our case, Sp = 1 because FP = 0. Although this concerns few cases (roughly 7.5% of all class1 patients), a specific test is always extremely valuable and rare in medicine. We have therefore proven the *existence* of such a test, unfortunately, we are not yet able to pinpoint which lab.exams were primarily responsible for such a classification, given the set of rules obtained from our algorithm. So this can be considered as a potentially very interesting result,

needing further research. Once identified, those lab.exams will help establish a diagnosis with absolute certainty (since the test is specific), and therefore will allow to start a treatment without further costly examinations. The interest is therefore both human (by limiting the number of tests necessary to reach a diagnosis), and economic (by avoiding unecessary and costly spc.exams).

5.3 A finding of medical interest

Taking the "unfolded" table of type II patients and retaining only the patients suffering from lupus (they test positive for CNS.lupus), the thrombosis class was predicted using the created "age of the disease" variable, and the ACL igM spc.exam. Once again, we used the 5*2 fold cross validation technique, and obtained an overall correct prediction rate of 98.82%, with the following confusion matrix:

	$class0$	$class1$	$class2$
$class0$	369	0	0
$class1$	7	35	0
$class2$	1	2	433

Note that each time the algorithm classified a patient in the class2 category, the prediction was always correct.
Then we tried to identify a new classification rule that was both medically exploitable, and using only lab.exams, to reach the same perfet prediction level for class2. We ended up with the following rule:

IF diagnosis=cns.lupus
AND IF DNA>5.57
THEN the patient is class2.

Medical interest: collagen diseases evolve by successive outbursts. During those outbursts, the body reacts by producing certain componants of the immune system. In the past years of the pKDD challenge, the correlation between a rise in the ACL igG rate and a thrombosis attack was established. What we have found here, in the first place, is a correlation between a rise in the ACL igM rate, and an outburst of cns.lupus, which is a particular type of collagen disease. The main interest of this result resides in the fact that igMs are the first immunoglobulins to be secreted when the immune system is stimulated (and for instance, in the case of rheumatoid polyarthritis which is another collagen disease, these are the only immunoglobulins to be secreted in the majority of cases), and in particular they are secreted prior to the igGs.
Moreover, our classification rule identifies the lab.exam which needs to be regularly checked to predict a future thrombosis attack with certainty, giving us also the alarm threshold.
Going back to our analogy, this means that if we know that the bridge is fragile (lupus) and we know which pillar to watch to have an idea of the state of fragility of the bridge (igMs), then by checking the level of a particular type of

traffic (DNAs), we will be able to predict the future collapse of the bridge with certainty (thrombosis attack).

6 Conclusion

The results obtained are surprising from several viewpoints. First of all, it is seen that spc.exams and lab.exams are sufficient to predict the health state with high | though not perfect |accuracy (99.28%). Further work will examine in more detail the misclassified cases, in order to understand, if possible, the causes of error.
Second, it is seen that in some particular contexts, spc.exams are not necessary to conclude on the risk of thrombosis. Third, for a particular type of collagen disease (lupus), one spc.exam (ACL igM rate) and one lab.exam (ADN) were seperately identified as good predictors of future thrombosis attacks. It is therefore possible to get an evaluation of the risk that a patient diagnosed with lupus has of developping thrombosis in the future, by testing igMs, and to check regularly his ADN level in comparison with the threshold earlier reported, to predict the advent of the attack.

On-going experiments focus on:
Learning on Type III patients for example (i.e generating a rule using only Spc.exams), and testing on Type II patients, is it possible to find a combination of rules on lab.exams, which will correctly classify the Type II patients for whom the learned rules do not hold? Is it possible to do that in general, or do we have to break this objective down into as many different tests as there are different collagen diseases?
The idea here is to start by segmenting all Type II patients according to whether the learned rule holds or not: a new column is therefore added to the description of Type II patients, taking the value 1 if the learned rule holds, 0 otherwise. Then, is it possible to predict this column, using Type II patients' lab.exams? Is this new set of rules concerning lab.exams medically exploitable?

In conclusion, it appears that medicine offers many challenging problems to data mining, which we all are interested in addressing for a better prevention of diseases and less invasive examinations.

References

[1] U.M. Fayyad and G.Piatetsky-Shapiro and P.Smith. (1996) From Data Mining to Knowledge Discovery: An Overview. *Advances in Knowledge Discovery and Data Mining, MIT Press*, 1-34

[2] Blum and Mitchell (1998). Combining Labeled and Unlabeled Data with Co-training *:COLT:Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*

[3] Quinlan J.R. C5.0 Data Mining Tool. www.rulequest.com, 1997

[4] Dietterich, T.G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms *: Neural computation, 10*, 1895-1923