DYNAMIC DATAMINING WITH COMPLEXITY THEORY – THE CASE OF PROGNOSING SEPSIS

Matej Mertik*, Miljenko Križmarič**, Zoran Zabavnik***, Milan Zorman*, Gregor Štiglic*, Dušanka Mičetič Turk**, Peter Kokol*

* University of Maribor, FERI, SI-2000
** University of Maribor, College of Nursing Studies, SI-2000
*** General Hospital Maribor, SI-2000

matej.mertik@uni-mb.si

Abstract: Sepsis is a severe illness caused by overwhelming infection of the bloodstream by toxin-producing bacteria. During the treatment of the sepsis we are facing example of self-organisation adaptation in the patient's bloodstream. In this paper we present an experiment of dynamic datamining where by each patient three different signals were sampled: number of leukocytes, c-reactive protein, and lactate. Based on the complexity theory, first some interesting characteristics of the signal were extracted: trend of the signal, type of the curve, fractal characteristic of the signal, and minimal and maximal deviation of the curve. The dataset of the signal's characteristics was later integrated with other anamnesis parameters into a dataset which was analysed with the Multimethod datamining tool (advanced datamining tool comparable with WEKA and Orange), which was developed in the Laboratory for System Design at the University of Maribor. In this preliminary study we got quite promising results, which can be helpful by predicting treatments of the sepsis. Therefore we have intent to enlarge our database with the additional parameters of the medicaments and more collected objects. In the future we are planning to build decision support system for sepsis treatment.

Introduction

In the medical dictionary the term "sepsis" is defined as the presence of various pus-forming and other pathogenic organisms or their toxins in the blood or tissues [1]. Generally sepsis represents a type of bloodstream infection that can be acquired by patients when they are in the hospital for another reason. If sepsis occurs by the patient, fast identifying of infection and appropriate treating is crucial.

The sepsis treatment process is time oriented. Within the process we are facing an example of self-organisation and adaptation of pathogenic organisms or their toxins in the blood. Looking on the treatment from the physical view, the sepsis can be observed like a dynamic process, from were we extracted some of the characteristics for sepsis prediction. The rest of the paper in formed as following. In the second section we present the experiment and collected data in prognosing the sepsis. We continue with description of dynamical characteristics, which were extracted for a datamining process. Further we describe Multimethod datamining tool, which we used for the experiments. At the end we show some of the interesting preliminary results and conclude the paper with future work.

Sepsis and important characteristics

The medical specialist and experts agree that important indexes prediction of the sepsis are number of leukocytes (1), volume of c-reactive protein (2), and volume of lactate (3) in the bloodstream. Therefore we captured the time signals based on the mentioned parameters during the sepsis treatment for each patient from the database, which was provided by General hospital Maribor [2].

The blood parameters evolved during the treatment according to the patient situation. The example of self-organisation and self-adaptation of pathogenic organisms or their toxins in the patient's blood has occurred. At the end of the treatment observed parameters formed the signals with the success of the patient treatment. For the dynamical datamining, such signals were analysed with some of the important characteristic. We calculate following descriptors: trend of the signal, type of the curve (signal), fractal characteristic of the signal (curve's fractal dimension), number of oscillations in the signal, and minimal and maximal deviation of the signal. In the table 1 the generated and anamnesis parameters for the sepsis database are numbered. Below on figure 1, an example of c-reactive parameter time signal is presented.

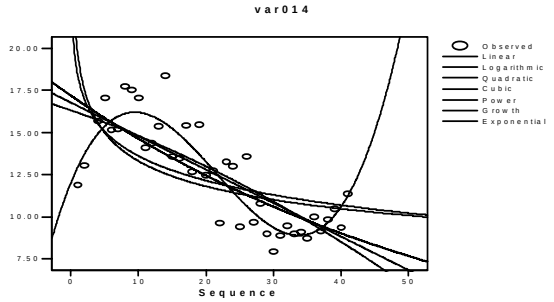
Table 1: generated (1) and general (2) parameters

Parameters/Signals	leukocytes	c-reactive	lactate
Trend	1	1	1
Oscillation	1	1	1
fractal dimension	1	1	1
Curve type	1	1	1

ISSN: 1727-1983

Age	2
Gender	2
Succes	2

Figure 1: c-reactive signal – curve approximation



The characteristic of the signal fractal dimension was calculated from the equatiation for power type curve approximation, where exponent b1 was considered as the complexcity measure.

$$Y = b_0 \times t^{b_1} \qquad (1)$$

Multimethod datamining tool

In the previous section we presented pre-processing of the dynamical data for the datamining process. We have seen on the table 1, the dataset with 15 different parameters generated from the signals was defined. In the dataset there were total 47 of objects (patients). 18 cases represented succeesful treatments and 29 un-succeesful treatments of the sepsis. In this section we describe the Multimethod datamining tool, which was applied to the generated dataset.

Multimethod is advanced datamining tool comparable with WEKA [3] and Orange [4] but with additional ability of knowledge transformation of various datamining methods in dynamic matter, and simple end-knowledge representation in a form of decision trees. For the experiment we used the arex [5] Multimethod algorithm within Multimethod tool, with which we have built decision trees on sepsis database for sepsis prognosing. We decided for the arex algorithm based on the experiments. By the experiments the comparison of the arex and standard method C5.0 [6] decision trees building based on the same dataset of sepsis showed up that the C5.0 achieved similar or even better accuracies of classifier than arex, however on the first and detailed view the trees generated with the standard C5.0 algorithm were to big, uninterested for the experts and they sloped to the overfitting.

Arex Multimethod algorithm obtains smaller sizes and deepness of the trees. In fact arex procedure combines different datamining methods (ID3 algorithm, Support Vector machines) with different ways of building trees (purity measures like chi quadrate, J measure…) within genetic environment. The optimal combination of such classifiers is generated trough the genetic evolution based on the best accuracies of the combined classifiers considering the dataset. The final combination of classifiers is combined in the final decision tree, which represent simple final end-knowledge representation for the users – medical experts. Readers, interesting in more detailed description of the Multimethod, which was developed in the Laboratory for System Design at University of Maribor, are encouraged to read [7,8].

Results

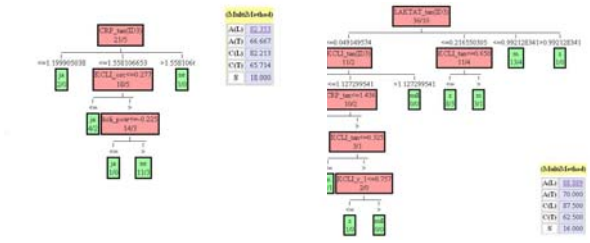
In this preliminary study of prognosing the sepsis we got quite promising results, which can be helpful for predicting treatments of the sepsis. Following, two examples of the final-decision trees are presented. In the first example the prediction for the treatment of sepsis based on the success attribute of the treatment, was made. In the second example the decision attribute gender of the patient was defined to be predicted.

Figure 2 present decision trees generated by Multimethod on the dataset of the sepsis. As we can see in the table 2, the accuracy of the generated trees with the arex algorithm for the both cases was between 80% and 90 % on the learning set and 65% and 80% on the test set, what represent good direction for prediction of the sepsis attributes.

Table 2: accuracies of classifier for two different examples

Decision attribute	Learning set	Test set
Example 1: Succes	88.88%	70.00 %
Example 2: Gender	82.35%	66.66%

Figure 2: extracted knowledge – Multimethod decision trees



In addition some of the interesting patterns - rules that were extracted from the trees by the experts are presented. As we can see on the figure 3, the rules for predicting the success of the sepsis treatment considered following parameters to be important: trend of the leukocytes, oscillation of the leukocytes, oscillation of lactate, and fractal dimension of the leukocytes. The parameters for prediction of the gender however considered trends of leukocytes, c-reactive and lactate.

ISSN: 1727-1983

Figure 3 presents some of the extracted rules by the experts.

Figure 3: Interesting patterns extracted form the results by the experts

Succees of the treatement [yes, no]

- **trend_leukocytes** <=0,367 AND **osc_lactat**=14 AND **pow_leukocytes**[-0,37, 0,92] THEN **no**
- **trend_leukocytes** >=0,69 AND **osc_lactat** <0.714 THEN **yes**
- **trend_leukocytes** [0.367 - 0,69] AND **osc_leukocytes** [0.13, 1.33] THEN **yes**

Gender [female, male]

- IF **trend_lactat** [0.049149574,0.216550305] AND **trend_leukocytes** <=0.650 THEN **female**
- IF **trend_lactat** [0.216550305 0.992128341] -> **male**
- IF **trend_lactat** [<=0.049149574] AND **trend_leukocytes** <=1.12729954 AND **trend_c-reactive** <= 1.436 THEN **male**

As we can see from the results above the prediction of the gender is accurate in approximately 83% and we can define the male for example on the trend of the lactate. We can see the interval on which the value of lactate trend is bounded; in this case the number interval is between 0.22 and 0.99.

Discussion and conclusion

Sepsis is a severe illness caused by overwhelming infection of the bloodstream by toxin-producing bacteria, which can be acquired during hospitalization of the patients. In this paper we showed the preliminary study of prognosis the sepsis based on the data which were gathered during the tree year treatment in general hospital in Maribor. For this purposes important attributes like number of leukocytes (1), volume of c-reactive protein (2), and volume of lactate (3) were extracted from the database during the treatment. The time signals of mentioned parameters, defined by the experts, were characterized with its main characteristic of the trend, number of oscillations, and also by their fractal dimensions. Such generated parameters with anamnesis parameters defined pre-processed dataset for the datamining process. Multimethod datamining tool with the advanced arex-multimethod algorithm was used for this purpose - a datamining tool comparable with WEKA and Orange developed in the Laboratory for System Design at the University of Maribor.

In preliminary study of prognosig the sepsis we got quite promising results, which can be helpful for predicting treatments of the sepsis. We showed two examples of the final-decision trees generated for prediction of the treatment of sepsis and prediction of the patient's gender; where in both cases the accuracy between 80% and 90 % on the learning set and 65% and 80 % on the test set was achieved. THis represents a good direction for prediction of the sepsis attribute. However at this moment we have only 47 objects in the dataset. Therefore intend to enlarge our database with more objects and add the additional parameters for the medicament's signals. Based on such future experiments and work, we are planning to build decision support system for the sepsis treatment.

References

[1] DIRCKX J. H. (1997): ‚Stedman's Concise Medical Dictionary for the Health Professions', 3rd edition, (Williams & Wilkins, Baltimore, Maryland), p. 792.

[2] MEDIS – Hospital Information System, General Hospital Maribor

[3] WITTEN I. H., FRANK E. (2005):‚Data Mining: Practical machine learning tools and techniques', 2nd Edition (Morgan Kaufmann, SanFrancisco).

[4] DEMSAR J, ZUPAN B, LEBAN G (2004): 'Orange: From Experimental Machine Learning to Interactive Data Mining', White Paper, (www.ailab.si/orange) Faculty of Computer and Information Science, University of Ljubljana.

[5] PODGORELEC V., KOKOL P. (2001): "Towards more optimal medical diagnosing with evolutionary algorithms", *J. med. syst., 2001*, vol. **25**, no. **3**, p. 195-219

[6] QUINLAN J.R. (1993): 'C4.5: Programs for Machine Learning, (Morgan Kaufmann publishers, San Mateo, CA).

[7] LENIC M. (2003): 'Multimetodna gradnja klasifikacijskih sistemov' – Phd. Thesis, University of Maribor.

[8] LENIC M., KOKOL P. (2002): 'Combining classifiers with Multimethod approach' - V. Second international conference on Hybrid Intelligent Systems, Santiago, Chile 2002.

ISSN: 1727-1983