Pattern discovery in intensive care data through sequence alignment of qualitative trends data : proof of concept on a diuresis data set

Martijn Devisscher MARTIJN.DEVISSCHER@UGENT.BE
Bernard De Baets BERNARD.DEBAETS@UGENT.BE
Ingmar Nopens INGMAR.NOPENS@UGENT.BE
Ghent University, Dept. Applied Mathematics, Biometrics and Process Control, Coupure Links 653, Ghent, Belgium

Johan Decruyenaere JOHAN.DECRUYENAERE@UGENT.BE
Dominique Benoit DOMINIQUE.BENOIT@UGENT.BE
Ghent University Hospital, Dept. Intensive Care Medicine, De Pintelaan 185, Ghent, Belgium

Keywords: qualitative data representation, monotone functions, k-means clustering, sequence alignment, intensive care, data mining

Abstract

This poster describes a methodology for mining intensive care data based on the application of a modified k-means clustering algorithm using sequence alignment. The methodology is applied to a urine production data set and is able to produce meaningful results.

1. Introduction

In recent years, automisation in intensive care units (ICUs) is expanding at a rapid pace. This trend results in a rapidly expanding high density data set, fostering possibilities for automated recognition of patterns, and possibly the discovery of new knowledge. This paper will focus on pattern discovery in time series data of an ICU.

In the described methodology, the raw data are first translated into a series of regularly sampled meaningful qualitative labels. This kind of abstraction is widespread in intelligent clinical data analysis (Stacey & McGregor, 2007) and guarantees that the obtained patterns are both easily interpretable by the medical experts and readily manipulated by the algorithms.

At every sampling instant, two labels are assigned. The first indicates the absolute level of the variable (e.g. Low, Medium, High), the second indicates the trend (e.g. Decreasing, Stable, Increasing). The inclusion of trend representations has proven useful in intensive care monitoring, e.g. for data validation (Horn et al., 1997) or for reducing false alarm rates (Charbonnier & Gentil, 2007). Furthermore, a nearly identical form of temporal abstraction has been used in prognostic prediction in ICU data (Verduijn et al., 2007).

The abstraction reduces the time series mining problem to that of mining a series of symbols, a problem well known in bioinformatics, where sequence alignment and clustering are widely used, for instance to identify functionally important regions in protein sequence data. Sequence alignment methods have already been succesfully applied to the analysis of qualitative trend data in the process industry (Balasko et al., 2006). The techniques used for sequence alignment are designed to deal with non-exact matching of strings, random insertions and deletions of symbols, features that are strongly desirable when comparing sequences of qualitative labels. Indeed, physiological reaction rates are highly inhomogeneous among individuals, and small labelling errors (stemming from measurement errors or from data processing) should be tolerated to some extent.

This poster describes a methodology for discovering patterns in time series data using a modified k-means clustering algorithm that uses sequence alignment to measure distance between instances. As a proof of concept, the methodology is applied to discover patterns in urine production data as a consequence of adminis-

Appearing in the Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland, 2008.

tering pharmaceuticals.

2. The data set

The data used consists of a set of urine production data for the first 3 days of stay of 1059 patients in the Intensive Care Unit of the Ghent University Hospital (Belgium), along with a list of the times, doses and id of administered pharmaceuticals.

Urine production data consist of series of time-volume measurements, corresponding with the volume of urine produced since the last emptying of the urine vessel. The time interval between measurements is variable.

3. Methodology

3.1. Data pretreatment

Before translation to qualitative labels, the data is pretreated to yield values for urine production rate and change in this rate at regularly spaced intervals. The interpolation procedure is as follows.

The discrete urine volume data points are cumulatively summed up, to generate a monotone urine volume curve.

A smooth, monotone function is fitted to these data (Ramsay, 1998). Although the method actually produces *strictly* monotone functions, monotone functions can be approximated arbitrarily well, by assuming a very small increase in the data over time. A fixed smoothness penalty λ value of 10 was used for all curves. Time was measured in hours. The technique was chosen because of its simplicity and because it easily provides estimates of first and second derivatives of the interpolated data.

The interpolation procedure therefore results in estimations at fixed time intervals of urine production rate (the first derivative of the cumulative urine production curve) and the rate of change in urine production rate (the second derivative).

This approach is specific for the case of urine production data because of its inherent monotonicity. However, any interpolation method that can also produce (smooth) estimations of derivatives (e.g. b-splines) can be used when dealing with non-monotone data.

An example of raw and pretreated data is given in figure 1.

3.2. Qualitative representation

For every sampling instant (taken every hour), the interpolated values are translated into qualitative labels

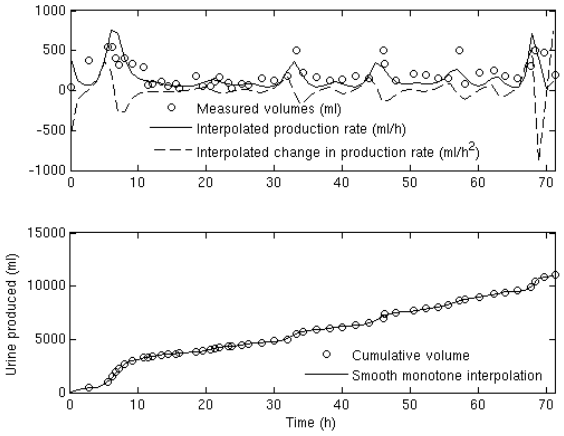

Figure 1. Illustration of the interpolation of irregularly measured urine volumes to regularly sampled production rate and change in production rate for one patient

representing the absolute level and the trend of the data.

Production rate level is labelled as {L(ow), N(ormal), H(igh)} and trend is labelled as {D(ecreasing), S(table), I(ncreasing)}. This labelling is performed by classifying the values of each of the two signals using fixed intervals.

As an example, the first 10 hours of the graph depicted in figure 1 are represented by

$$\begin{array}{cccccccccc} H & N & N & N & N & H & H & H & H & H \\ D & D & S & S & I & I & I & D & D & D \end{array}$$

The data are then subdivided into substrings that are meaningful in the current problem setting. In this proof of concept, substrings were selected that contain a fixed time interval after administration of a certain pharmaceutical.

3.3. Pattern discovery using clustering and sequence alignment methods

K-means algorithms have been successfully applied for finding patterns in protein sequences (Zhong et al., 2005). We use a modified k-means algorithm to identify clusters in the substring collection in order to find characteristic patterns.

For each cluster, a consensus pattern is constructed using majority voting on each label position. This means that, for every position in the string, the most frequently occurring label is selected to represent the cluster. The consensus pattern then functions as the

centroid of the cluster. To compare substrings with the centroid, the complement of the similarity score of a global optimal alignment of the two substrings is used as a distance measure.

The alignment is performed using a dynamic programming algorithm based on the Needleman-Wunsch algorithm (Gusfield, 1997). Essential to this algorithm is a similarity matrix, which is a matrix quantifying similarity between any two symbols in the substring. The algorithm was originally designed for single symbol strings, but extension to multisymbol strings, such as the two-symbol string in this study, is straightforward, since each combination of symbols can be considered a single new symbol. However, such an approach leads to an explosion of the similarity matrix. Therefore, instead of constructing a 9x9 similarity matrix representing similarity between any two combinations of the labels on the two signals (production rate and rate change), we neglected cross-dependencies between the signals and used the sum of similarities within each signal.

The gap penalty was set to 0, and identity matrices were used as similarity matrices.

4. Results

The methodology was applied to substrings representing the trend in urine production for six hours after administration of the most frequently administered drug. This query resulted in a subset of 700 strings of 6 sampling points, each represented by 2 labels from the subsets {L(ow), N(ormal), H(igh)} and {D(ecreasing), S(table), I(ncreasing)}.

Running the algorithm with two classes, the following class centroids were identified:

$$\begin{matrix} N & H & H & N & N & N \\ I & I & D & D & D & S \end{matrix}$$

for the first class, in 241 of the cases, and

$$\begin{matrix} N & N & N & N & N & N \\ S & S & S & S & S & S \end{matrix}$$

for the second class, in 459 of the cases.

These are meaningful patterns that can be easily interpreted as follows: the administration of the pharmaceutical either has no effect at all, or results in a temporary increase in production rate to a high level for about two hours, to fall back to normal afterwards. After further investigation, the pharmaceutical appeared to be furosemide, which indeed acutely stimulates diuresis for a few hours.

The results are somewhat dependent from the initialisation of the k-means algorithm, since our current method uses random initialisation. However, since it is our goal to extract meaningful patterns, the clustering can be repeated a number of times with an increasing number of classes until no new patterns are discovered.

5. Conclusions and Perspectives

The results indicate that the methodology described here is able to extract meaningful patterns, and is therefore a promising tool for mining ICU data. The technique will now be applied to multivariable datasets, where the selected substrings are selected within a time frame before certain adverse events in ICU practice. The ultimate goal is to use the discovered patterns in alignment algorithms for automated detection of conditions leading to these adverse events.

References

Balasko, B., Nemeth, S., & Abonyi, J. (2006). Qualitative analysis of segmented time-series by sequence alignment. *Proceedings 7^{th} International Symposium of Hungarian Researchers on Computational Intelligence.*

Charbonnier, S., & Gentil, S. (2007). A trend-based alarm system to improve patient monitoring in intensive care units. *Control Engineering Practice*, *15*, 1039–1050.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences.* Cambridge University Press, New York, USA.

Horn, W., Miksch, S., Egghart, G., Popow, C., & Paky, F. (1997). Effective data validation of high-frequency data: time-point-, time-interval-, and trend-based methods. *Comput. Biol. Met.*, *27*, 389–409.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *J. R. Statist. Soc. B*, *60*, 365–375.

Stacey, M., & McGregor, C. (2007). Temporal abstraction in intelligent clinical data analysis: a survey. *Artificial Intelligence in Medicine*, 1–24.

Verduijn, M., Sacchi, L., Peek, N., Bellazzi, R., Jonge, E., & de Mol, B. (2007). Temporal abstraction for feature extraction: A comparative case study in prediction from intensive care monitoring data. *Artificial Intelligence in Medicine*, 1–12.

Zhong, W., Altun, G., Harrison, R., Tai, P., & Pan, Y. (2005). Improved k-means clustering algorithm for

exploring local protein sequence motifs representing common structural property. *IEEE Transactions on Nanobioscience*, *4*, 255–265.