Multivariate Time Series Analysis of Clinical and Physiological Data

Patricia Ordóñez Rozo
UMBC
1000 Hilltop Circle
Baltimore, MD 21250
patti.ordonez@umbc.edu

Marie desJardins
UMBC
1000 Hilltop Circle
Baltimore, MD 21250
{mariedj}@cs.umbc.edu

Abstract

We aim to create a multivariate temporal representation of electronic medical data which automates the personalization of baselines and thresholds based on a patient's history. Visualizations based on the representation emphasize the rate of change in variables and assist providers in analyzing the data from a multivariate perspective. A novel similarity metric for this representation will be the cornerstone to the development of a search engine for large medical databases.

1. Problem Description

Although the sophistication and volume of collected data is greater now than at any point in the history of medicine, the information overload that providers face may inhibit the diagnostic process [5]. Providers are expected to examine large volumes of data and identify correlations between dozens to hundreds of parameters based on their own clinical experience in order to detect significant medical events. Yet, a psychological study in 2005 found that humans can process up to four independent variables in bar graphs of statistical data accurately and efficiently. When a fifth variable was introduced, accuracy dropped to that of a coin toss [4]. Existing visualizations of the data to assist the provider in analyzing the information consist of a table or plot of values for a particular parameter over time. Thus, we propose multivariate visualizations to assist providers with the analyzing of the large amounts of data they are provided. These visualizations will not only capture the relationships between the variables, but also the rate of change of the variables over time.

Automated techniques for discovering the correlations between parameters would assist the provider in making a diagnosis and would help in identifying hidden patterns within the data associated with specific medical conditions or events. Current techniques focus on querying these databases using conjunctions and/or disjunctions on ranges of different parameters.

Furthermore, these queries are often based on traditional baselines and thresholds for parameters, which indicate "normal" ranges of values in a healthy patient. These values are generally based on age, gender, and/or weight, and were established (often decades ago) by measuring these parameters in a large population of patients. While these values are appropriate for many people, they are not for others because of differences in lifestyle, genetic makeup, and environment. With the development of electronic medical records, it is now possible to use a patient's medical history to automate the personalization of a patient's baselines and thresholds.

Thus, we aim to create a search engine such that providers will be able to enter a patient's data and retrieve "medically similar" patients placing more emphasis on particular parameters if desired and taking into account a personalized representation of each patient's data.

2. Proposed Research

Hypothesis

We hypothesize that it may be possible to:

- Create a visualization that will assist providers in examining multivariate patient data over time more accurately and efficiently than current tabular visualizations,
- Identify hidden patterns in medical data that would signal significant medical events (such as organ failure) hours in advance, and
- Develop a measure of similarity for multivariate time series representations of physiological and clinical electronic data allowing physicians to identify patients with similar events and/or phenotypes for the purpose of predicting patient outcomes.

Background

Much of the previous research on time series analysis has focused on univariate time series data. An extensive survey of the history of times series research is given by Keogh [6].

In order to simplify data mining for the time series data, we considered using several alternative time series representations, including Discrete Fourier Transformation [1], Discrete Wavelet Transformation [3], and Piecewise Linear Approximation [6]. We chose to use

SAX [7] because of its ease of computation and comprehension, its ability to reduce the data's dimensionality, and to provide a lower bound on Euclidean distances between time series. The latter characteristic allows the use of data mining algorithms on the symbolic representation that they give identical results as on the original time series.

SAX also is well known for its ability to detect motifs and anomalies in univariate time series data. This aspect is particularly appealing because we expect significant medical events to appear as anomalies and the phenotypic profile of a person's data to appear as a collection of motifs.

Prior to discretizing the univariate time series into SAX, the data is normalized such that the average is 0 and the standard deviation is 1. Thus, the data is compared to itself and personalized.

SAX then divides the time series *C* (shown as a red curved line in Figure 2) of length *n* into *w* segments of equal width along the *x*-axis. The time series in Figure 1 has $w = 8$. The average of the values within each segment is used as the value for the segment in a new discretized time series. This representation of time series is known as Piecewise Aggregate Approximation (PAA, represented by the blue, green, and aqua colored horizontal lines in Figure 1). SAX divides the time series into an equal distribution of parts, depending on the desired alphabet size. It assumes a Gaussian distribution of values, so for an alphabet size of 3, the separations occur at -.43 and .43, as shown by the straight horizontal gray lines in Figure 1. The PAA value for each segment is then converted into a symbol, depending on the part of the distribution into which the PAA value falls (labels **a**, **b** and **c**).

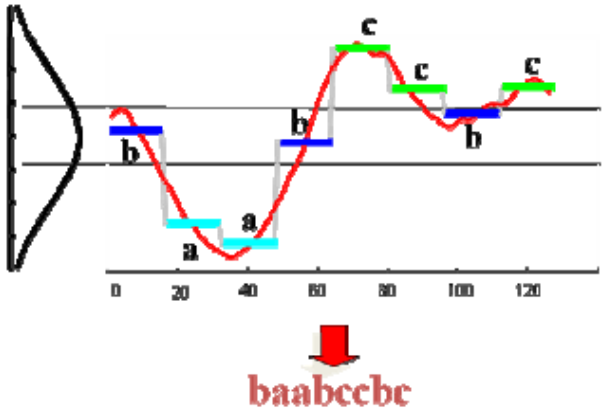

Figure 1: SAX as represented by Keogh [6]

The problem with using SAX for our purposes is that SAX requires univariate time series of equal length, and in our case, we will be examining multiple time series per individual and each time series may be of a different length and the ranges for each parameter are orders of magnitude different.

Thus, we are considering using Bag-of-Patterns (BOP), which is based on SAX and focuses on comparing the structural similarity of univariate time series. BOPhas been found to outperform existing techniques for classification, clustering and anomaly detection of univariate time series [8].

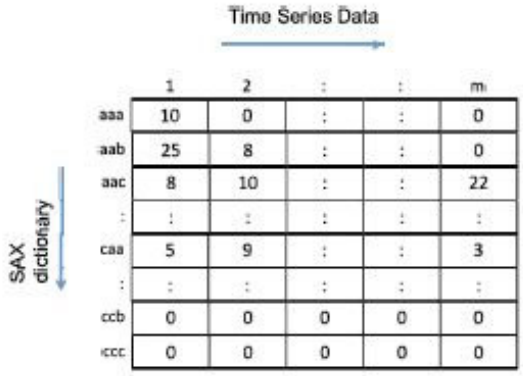

Figure 2: BOP representations of *m* time series as represented by Lin [8]

Bag-of Patterns (BOP) is based on the Vector Space Model [9] (a.k.a."bag of words" model) for comparing similar documents. In the Vector Space Model, documents are considered similar if they have a large number of unusual words in common. Frequent trivial words such as "the" and "a" are ignored and the juxtaposition of the words is not considered. With BOP, two time series are similar if they have a large number of uncommon time series subsequences in common.

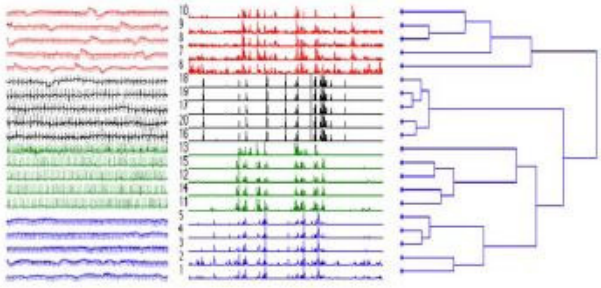

Figure 3: The histogram of the BOP representation of 20 time series and the resulting dendrogram after hierarchical clustering as represented by Lin [8]

In BOP, each time series is converted into a series of words of length *n*. These words are calculated by having a sliding window of length *n* over the entire time series and converting these subsequences into a SAX representation. The frequency of the words is stored in a vector. Identical consecutive words are only counted once to avoid matching based on recurring patterns (like trivial words in a document). Figure 2 displays the frequency vectors for a

collection of *m* time series. Distance between the vectors can be calculated using any conventional distance metric.

An example of using BOP in hierarchical clustering is shown in Figure 3. Each time series is followed by a histogram of its frequency vector. Euclidean distance was used as the similarity metric. The resulting dendrogram is on the right. Although BOP focuses on comparing time series based on structure rather than the shape, it appears to cluster time series better than using Euclidean distance on shape-based representations such as the original series, Dynamic Time Warping, and Discrete Fourier Transform.

3. Approach

Our approach is to build a prototype of a visualization tool and search engine. The majority of the research completed thus far is with the visualization tool. We are beginning to break ground on the implementation of the similarity metric. The search engine will be developed once the similarity metric has proven to be effective. This section is broken into three parts: one to describe the work I have completed for the visualization tool, the work I am currently working on for the similarity metric, and the work I intend to do for the search engine.

Visualization

The visualization must enable providers to efficiently and accurately review a patient's physiological and clinical data from a multivariate perspective. The star plot [2], or radar diagram, is known for its ability to display multiple variables in a small area; however, the star plot is not designed for time series. Thus, I have created a visualization, named a Multivariate Time Series Amalgam (MTSA) Visualization [11], that captures the changes in a patient's state over time by overlaying star plots as seen in Figure 4. The color of the star plot is used to indicate temporal changes with the darkest color representing the most recent interval.

Each star plot represents the patient's state over that interval. The red dotted circle in the middle represents the average value of the patient's data over the entire interval for each parameter. Each point is plotting the standard deviation of the value of the interval. Thus, the first circle outside the red circle represents one positive standard deviation and the first circle inside one negative. The minimum and maximum values for the patient over the total time are indicated in square brackets beside the parameter name. A provider can review the values that were used to calculate a point, as well as the method that was used to calculate it (average or linear interpolation) by clicking on it.

The parameters are ordered according to their effect on one of four vital organs; the heart, liver, lung, and kidney, and then alphabetically. Each organ is represented by a color and the axes are colored accordingly. Functionality exists to change the colors associated with an organ as well as to view only the parameters for a particular organ by checking the check box next to the organ's button.

By clicking on the Start button, a provider can also view an animation of changes in the patient's state over time. Each star plot is highlighted over time starting with the earliest interval to the latest interval. Figure 5 shows the highlighting of the star plot at a particular time interval.

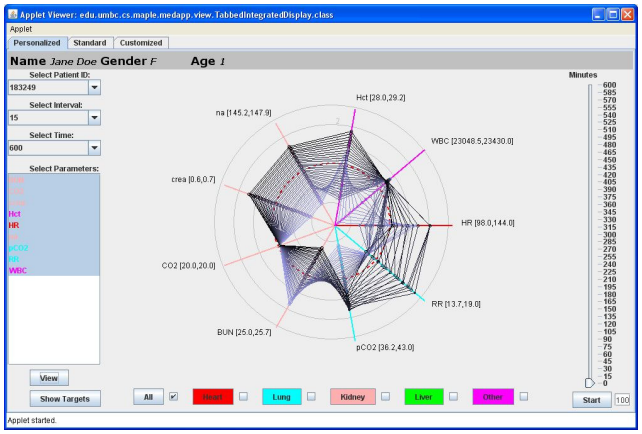

Figure 4: MTSA Visualization of 10 hours of data averaged in 15 minute intervals over 600 minutes

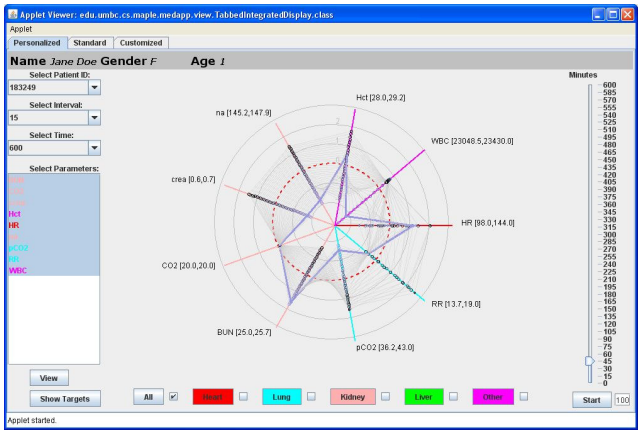

Figure 5: Personalized visualization highlighting the 30 to 45 minute interval

The provider may also toggle between two other views of the data by clicking on the Standard and Customized tabs in the upper left. The Standard view plots the values along each axis relative to a target value entered by the provider. The center point of the circle becomes equivalent to 0.

The Customized view allows a provider to enter a target as well as the step value for each parameter. This capability enables the provider to tailor the visualization in order to highlight the patient's "normal" values, represented by the yellow circle as well as define an acceptable upper and lower bound, represented by the two red dotted circles seen in Figure 6.

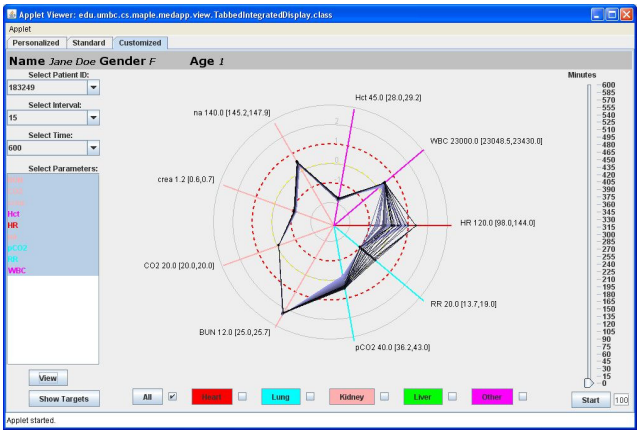

Figure 6: Customized view of patient in Figure 4

Similarity Metric

An underlying belief of our research is that the complexity of a human being is best captured by examining multiple data sets from the patient over time. We hypothesize that the organs of a human work intricately together to keep an individual alive, and thus hidden patterns in the medical data must exist if we examine the data from a multivariate perspective. Thus, we propose to create a multivariate version of the Bag-of-Patterns representation which captures the frequency of motifs and anomalies in the univariate time series from multiple sources to develop a similarity metric that can find medically similar patients.

We are in the beginning stages of this research and are examining a data set of 105 neo-natal patients from the neo-natal intensive care unit at Johns Hopkins hospital. The research will first focus on the classification of these patients into three categories; those with severe patent ductus arteriosus (PDA) that require surgery, those with slight PDA that will respond to pharmacological therapy and those with no PDA. Patent ductus arteriosus results in the lack of closure in the ductus arteriosus after birth and may result in death if left untreated. Current practices use electrocardiography as the gold standard for detecting PDA. However, since surgical intervention requires transfer to another facility, it is important to differentiate between those cases of PDA that require surgery and those which do not.

We will first be using supervised learning techniques to detect whether or not the hidden patterns can be detected in the data. Afterwards, we will focus on developing an unsupervised technique based on a similarity metric for our multivariate Bag-of-Patterns representation to cluster the data into categories. An unsupervised technique is crucial to the development of the search engine since labels cannot be provided for all medical events.

Search Engine

Once the similarity metric has been developed, classic information retrieval (IR) methods used for document similarity may be used. Classic IR methods estimate the similarity between two documents by determining the frequency of unusual words between them. A term frequency vector (TFV) is used to capture the frequency of the terms, as in the Vector Space Model. The words are weighted using an inverse document frequency (IDF) scheme, so that words that are less frequently observed are more heavily weighted [14]. To apply these concepts to the patient similarity problem, these IR methods must be modified to incorporate the finding of similar anomalies and motifs in patients' data taking into account the personalized representation of the data as well as the standard baselines and thresholds that currently exist for determining what an anomaly is.

4. Evaluation

Thus far, we have completed a pilot study on the three MTSA visualizations; Personalized, Standard and Customize. In the study, 14 internal medical residents at St. Agnes Hospital in Baltimore were presented with up to 10 hours of clinical data for 10 patients and were asked to predict whether the patient would experience an episode of acute hypotension in the hour following the data. The data came from the 2009 PhysioNet Challenge (http://physionet.org/challenge/2009/). Each resident used the visualizations for five patients and univariate plots of variable vs. time with the supporting table data for the remaining five.

Results indicated that the residents were 52% accurate using the visualizations and 56% accurate using the traditional method. Although neither group did much better than a coin toss, the residents had only been using the visualizations for over an hour whereas they had been using the tables for over 9 months; therefore, we see tremendous potential in the use of the MTSA visualizations. Interestingly, all of the automated methods submitted for the PhysioNet Challenge in 2009 correctly categorized the ten patients, reflecting the need for decision support systems in medicine.

Once the visualization tool and search engine are completed, we intend to do a usability study with providers at Johns Hopkins University. We have applied for a Collaborative Research Experience for Undergraduates grant for assistance in this endeavor.

5. Related Work

Current visualizations for electronic medical data in bedside monitors consist of stacks of univariate time series and/or tables of provider-validated data. The author is not aware of new visualizations for electronic medical data designed for the bedside monitor. Other existing medical visualizations for physiological and clinical data have focused instead on capturing the personal medical history of a patient over a lifetime.

Lifelines [12] is a visual summary of a patient's medical history. The data is converted to categorical data, meaning that significant medical events are grouped into categories, and are stacked as seen in Figure 7.

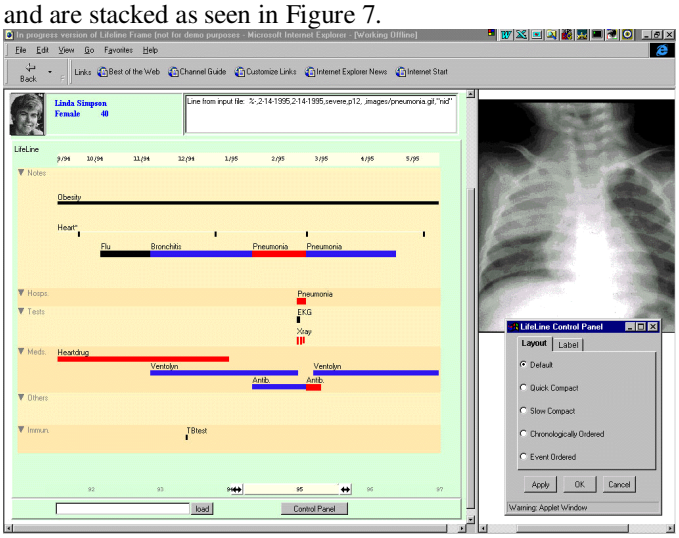

Figure 7:Lifelines visualization of personal medical history of a patient

Knave II [16] takes a similar approach to viewing a patient's medical history with the difference that the data is categorized with the intention of measuring a predefined state, like the platelet or liver state as seen in Figure 8.

The most similar visualization that has been created for multivariate time series data is Zoom Star [10]. Zoom Star overlays star plots in a 2D space with different colors and levels of transparency as seen in Figure 9, but does not include the animation of our visualization. Also, Zoom Stars are intended for interval data where the values are represented by the thickness of the colored lines along an axis or the size of a node along an axis. Figure 9 is displaying the range of values for eight stocks over three weeks. Our visualization plots time series data, not interval data.

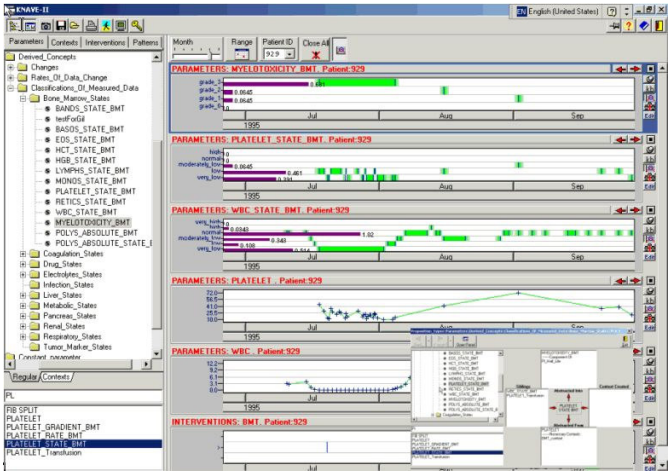

Figure 8: Knave II visualization of personal medical history of a patient

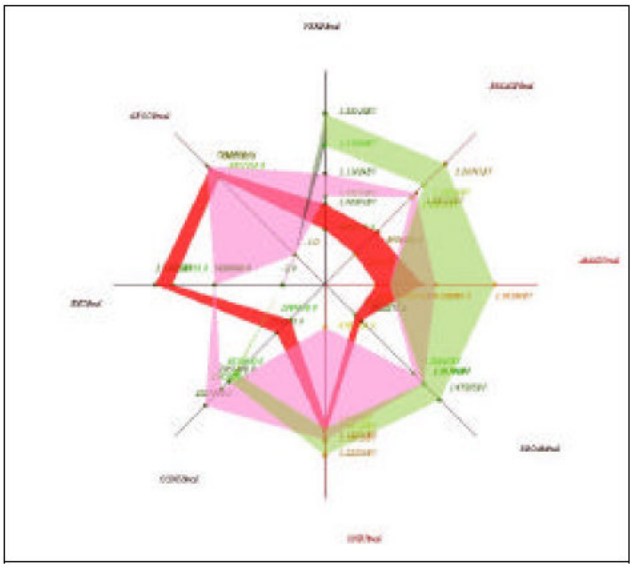

Figure 9: Zoom Star visualization of 8 stocks over 3 weeks

In the last few years with the emphasis on converting medical records into electronic health records, there has been an emphasis on developing data mining techniques for physiological and clinical data. We are highlighting only the work which inspired our work.

Saeed & Mark [13] used data mining techniques and heart rate, blood pressure and cardiac output measurements to determine whether similar patterns in patient's physiological data prior to hemodynamic deterioration could be indicative of an episode of severe hypotension. Their method converts the time series into Discrete Wavelets Transformations, and then uses the coefficients to create a frequency vector. A novel similarity metric, which emulates TF/IDF, and a k-nearest-neighbor classifier is used to create a predictor for hemodynamic deterioration. Their algorithm is a supervised learning algorithm and used only 3 parameters.

Sorani et al. [17] performed hierarchical clustering on 23 patients using 20 physiological parameters from the ICU. The data was captured every minute, a much higher sampling frequency than is available for our patients. As a result, no interpolation of the data was required and missing or unrealistic values were ignored. The data was normalized around patient and variable medians and then clustered using average linkage hierarchical clustering. The authors found that using a multivariate, highly sampled data set, they were able to create patient profiles for diagnosis and treatment.

6. Contributions

I expect to create a prototype of a novel and functional clinical decision support system (CDSS) by the time I complete my dissertation. The CDSS will include a

visualization tool and a search engine for a large electronic medical database.

7. Research Philosophy and Advice

My approach to research is to extract as many ideas as possible from solutions in other problem domains that have similar characteristics to my problem domain. I try to envision how the successful approaches of others can apply to my problem and allow them to inspire me to develop creative solutions. If you are tackling a big problem, break it down into smaller ones and focus on one problem at a time. When evaluating your research, focus on extracting meaning information from your results instead of proving your hypothesis is correct.

My greatest piece of advice is for women not to give up on a career path or problem simply because it is hard, or because people tell you, you cannot do it. Instead work harder. Find allies who can help you, advisors who offer wisdom and positive encouragement, and advocates who can help you to succeed. Challenge yourself to do things you never imagined because you will never know what you are capable of doing unless you push yourself beyond what is comfortable. Lastly, remember to give back.

Acknowledgments

The primary author would like to thank all her mentees, mentors, allies, advisors, and advocates. This work is funded by a National Science Foundation Graduate Research Fellowship.

Bibliography

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In Proc. of Foundations of Data Organization and Algorithms, pages 69-89, 1993.

[2] J. M. Chambers, W. S. Cleveland, B. Kliener, and P. A. Tukey. Graphical Methods for Data Analysis. Boston: Wadsworth International, Duxbury Press, 1983.

[3] K. Chan, A. Fu. Efficient time series matching by wavelet.In Proc. of ICDE, pages 126-133, 1999.

[4] G. S. Halford, R. Baker, J.E. McCredden, J. D. Bain. How many variables can humans process? In Psycological Science, pages 70-6, Jan 2005.

[5] T. Heldt, B. Long, G.C. Verghese, P. Szolovits, and R. G. Mark. Integrating data, models, and reasoning in critical care. In Engineering in Medicine and Biology Society, pages 350–353, 2006.

[6] E. Keogh. A decade of progress in indexing and mining large time series databases. VLDB 2006, http://www.cs.ucr.edu/~eamonn/tutorials.html, 2006.

[7] J. Lin, E. Keogh, L. Wei, S. Leonardi. Experiencing SAX: a novel symbolic representation of time series. In Data Min Knowl Disc, pages 107-144, 2007.

[8] J. Lin and Y. Li. Finding structural similarity in time series data using Bag-of-Patterns representation, In Proc. of SSDBM, pages 461–477, 2009.

[9] Y. Morinaka, M. Yoshikawa, T. Amagasa, and S. Uemura. The L-index: An indexing structure for efficient subsequence matching in time sequence databases. In PAKDD, pages 51-60, 2001.

[10] M. Noirhomme-Fraiture, Visualization of Large Data Sets: the Zoom Star Solution. International Electronic Journal of Symbolic Data Analysis, 2002.

[11] P. Ordóñez, M. desJardins, C. Feltes, C. U. Lehmann, J. Fackler. Visualizing multivariate time series data to detect specific medical conditions. In Proc. of AMIA Annu Symp, pages 530-534, 2008.

[12] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: Using visualization to enhance navigation and analysis of patient records. In Proc. of AMIA Annu Symp, pages 76–80, 1998.

[13] M. Saeed, R. Mark. A novel method for the efficient retrieval of similar multiparamenter physiologic time series using wavelet-based symbolic representations. In Proc. of AMIA Annu Symp, pages 679-683, 2006.

[14] G. Salton, E. A. Fox, and H. Wu. Extended Boolean information retrieval. Communications of the ACM 26(11), pages 1022-1036, 1983.

[15] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. Communications of the ACM 19(11), pages 613–620, 1975.

[16] Y. Shahar, D. Goren-Bar, D. Boaz, and G. Tahan. Knave-II: A distributed architecture for interactive visualization and intelligent exploration of time-oriented clinical data. In Proc. of AMIA Annu Symp, pages 115–35, 2006.

[17] M.D. Sorani, J.C. Hemphill III, D. Morabito, G. Rosenthal, G.T. Manley. New Approaches to Physiological Informatics in Neurocritical Care. Neurocritical Care, pages 45-52, 2007.