Visualizing Multivariate Time Series Data to Detect Specific Medical Conditions

Patricia Ordóñez[1], Marie desJardins, PhD[1], Carolyn Feltes, MD, PhD[2], Christoph U. Lehmann, MD[2] and James Fackler, MD[2]

[1]University of Maryland, Baltimore County, Baltimore, MD and [2]Johns Hopkins University School of Medicine, Baltimore, MD

Abstract

Efficient unsupervised algorithms for the detection of patterns in time series data, often called motifs, have been used in many applications, such as identifying words in different languages, detecting anomalies in ECG readings, and finding similarities between images. We present a process that creates a personalized multivariate time series representation —a Multivariate Time Series Amalgam (MTSA) — of physiological data and laboratory results that physicians can visually interpret. We then apply a technique that has demonstrated success with the interpretation of univariate data, named Symbolic Aggregate Approximation (SAX), to visualize patterns in the MTSAs that may differentiate between medical conditions such as renal and respiratory failure.

Introduction

The medical domain has long been an area where computer science has the potential to play a pivotal role. A national push for the use of electronic medical records[1] has not yet shown great success. Nevertheless, EMRs, when fully implemented, can store an immense amount of information about a patient's vital signs, laboratory results, medications and device data. The technology now exists to capture up to 350 different types of vital signs and laboratory reports on a patient in intensive care. However, this data is under-utilized: better tools are needed for multivariate time series analysis and visualization.

Current methods used to understand clinical data primarily focus on analyzing the data along a single dimension. For example, if a provider is reviewing the status of a patient with respiratory distress, she would examine the individual vital signs and laboratory results that pertain to the respiratory system. These vital signs and laboratory results each represent a univariate time series. The detection of an abnormal or unusual circumstance is typically based on the examination of the univariate parameters to identify values that fall above or below a preset threshold. However, because of parameter dependence and variation over time in a complex organism such as a human, examining the vital signs and laboratory results together in a multivariate time series may provide greater insight into how the body and its vital organs function as a whole and a means by which to better diagnose and treat a patient.

We present a process that interleaves univariate time series data into a multivariate time series representation, which we refer to as a *Multivariate Time Series Amalgam* (MTSA). We present a visualization of the resulting MTSA that groups related vital signs and laboratory results together and that displays the changes in each over time. We illustrate this visualization using the data of one critically ill child. We then create the MTSAs of the vital signs and laboratory data from five critically ill children. We show how to convert each MTSA into the SAX string representation that was originally created for use with univariate time series data.[2] Finally, we discuss how the resulting string-based representation could be used to group patients according to patterns in their physiological data.

Background

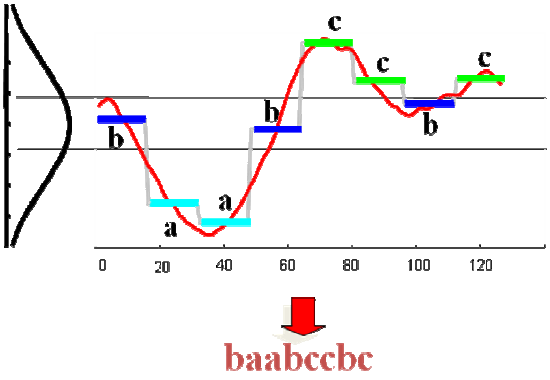
Much of the previous research on time series analysis has focused on univariate time series data. An extensive survey of the history of times series research is given by Keogh.[3]

In order to simplify the time series for data mining, we considered using several alternative time series representations, including Discrete Fourier Transformation,[4] Discrete Wavelet Transformation,[5] and Piecewise Linear Approximation.[6] We chose to use SAX[7] because of its ease of computation and comprehension and its ability to reduce the data's dimensionality and to provide a lower bound on Euclidean distances between time series. However, SAX is designed for univariate data. We demonstrate that MTSAs can be converted into a SAX representation with the goal of analyzing them effectively using methods that are normally reserved for univariate time series in future work.

SAX is well known for its ability to detect motifs and anomalies in univariate time-series data. It has also been used to measure the similarity of images and motions.

SAX first normalizes the data so that each time series has a mean of 0 and a standard of deviation of 1, assuming a Gaussian distribution of the values. SAX then divides the time series *C* (shown as a curved line in Figure 1) 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 colored horizontal lines in Figure 1). SAX then examines the Gaussian distribution of values and divides the distribution into equal parts, depending on the desired alphabet size. For example, for an alphabet size of 3, the separations occur at -.43 and .43 in a normalized Gaussian distribution, 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 section of the distribution into which the PAA value falls (labels **a**, **b** and **c** in Figure 1).

Figure 1: SAX diagram by Keogh[3]



Methods

Preprocessing of Data:

The quality of medical data is far from ideal because there are so many human factors that influence the measurement and documentation of the data. Therefore, significant effort is required to preprocess the data.

In the case of the Pediatric Intensive Care Unit (PICU) data used in this paper, several of the parameter values contained cancelled or missing data, meaning that the measurement was not taken at its designated time. These data points were ignored.

In addition, the measurements for the parameters varied in frequency from once every 15 minutes to once every few days. Since existing time series methods generally assume a constant sampling rate for all measurements, we used linear interpolation methods across the missing data in order to create the MTSA.

The physiological variables selected had to be easily and frequently measurable for any critically ill child. The criterion for the selected patients was that they experienced either respiratory or renal failure.

The selected parameters were separated into four categories, those associated with the cardiac, pulmonary, renal and miscellaneous systems. The order of the parameters can be seen in **Table 1**. In future work, we aim to increase the number of parameters for the cardiac and pulmonary categories. Here we were limited by the number of parameters that were common to all five patients.

Table 1: List of Physiological Parameters

Parameter	Category
Heart Rate (HR)	Cardiac
Respiratory Rate (RR)	Pulmonary
pCO2	
CO2	Renal
Creatinine	
BUN	
Na	
White Blood Cell Count (WBC)	Miscellaneous
Core Temperature	
Hematocrit (Hct)	

Modified PAA:

The difficulty in performing PAA on several univariate time series using the frame length *w* was the result of the missing data. Some of the vital signs and laboratory results were not recorded within an interval of the desired frame length, and all of the univariate time series needed to have the same window size in order to create the MTSA. To avoid

extrapolation and to evaluate all parameter values, we used the latest recorded parameter starting value and the earliest recorded parameter ending value. The times of these values became the boundaries for all of the univariate parameter time series for a given patient.

PAA was used for all frames that contained data. However, if a frame did not have data, the value for the frame was calculated by using linear regression between the points nearest the boundaries of the frame. In Figure 2, frames 1, 2 and 4 use PAA, and frame 3 uses linear interpolation.

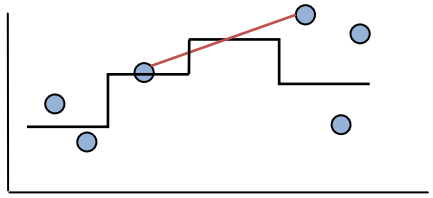


Figure 2: Linear Regression for Missing Values

Multivariate Time Series Amalgam (MTSA):

Once the univariate time series for a given patient have been calculated using the modified PAA, the MTSA can be created.

The primary objective of the MTSA was to create a multivariate representation of the univariate data that we could use to both analyze a patient's data relative to itself and to examine the patient as a whole, meaning that the representation focused on changes in the parameters and examined all of the parameters at once for a given instance. The second objective was to create a visualization that providers would find intuitive and inclusive and that would emphasize the changes in a patient's state over time.

In order to meet the first objective, the MTSA interleaves all of the values for each instance in a consistent order. To meet the second objective, the order for the interleaving was determined to be relative to the organs for which the vital sign or lab report provides information. Thus, all parameters that measure the state of a given vital organ are grouped together.

The MTSA Visualization:

Once the values for the individual time series have been calculated, then the multivariate time series is created by interleaving the values for an instance in time. A radial representation for the series at one instance is shown in Figure 3.

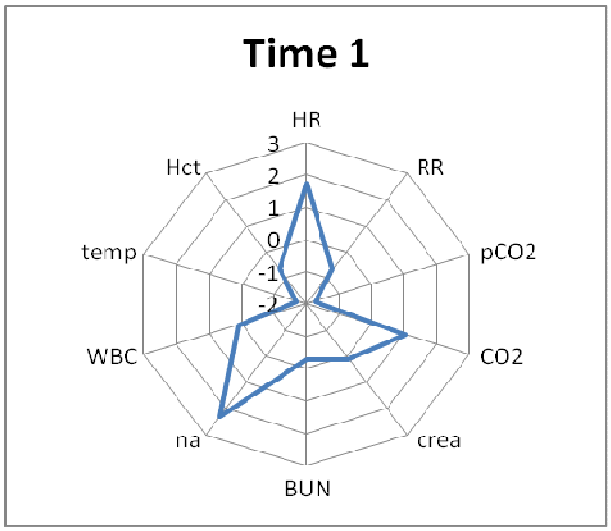


Figure 3: A radial representation of a multivariate time series for one patient at one instance in time.

The MTSA consists of an amalgamation of these instances. Several instances can be overlaid one another, as in Figure 4, to reveal a patient's progression over time for multiple variables. The temporal relationship of these overlaid multivariate time series is indicated by the color of the lines. The darkest line represents the most current instance in the MTSA and the lightest represents the oldest one.

From MTSA to SAX:

The resulting MTSA is a collection of pictures in time of a person's physiological data that are grouped in a manner to visualize more clearly the physiological changes in patient relating to one major organ. The MTSA can also be viewed as a simple time series; five MTSAs are displayed in this manner in Figure 5.

We cut the MTSAs for each patient into the same length, aligning them as advised by domain experts. On examination they look as if they are overlaid because of the recurring patterns.

Notice that some of the MTSAs such as the ones for patients 3 and 4 show a pattern of convergence at a central region. We are working on developing methods to find these patterns and discover their medical significance.

One of the methods we are currently using is SAX. We applied SAX to convert the MTSAs into string-based representations. We used the MINDIST[2] method to measure the distances between the strings and clustered them using a single link clustering method.

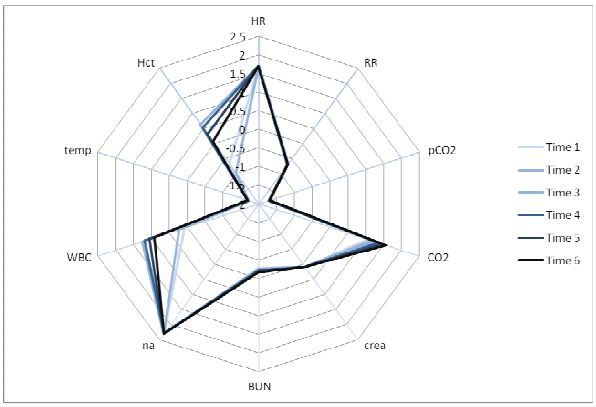


Figure 4: A radial representation of a multivariate time series for one patient over six hours.

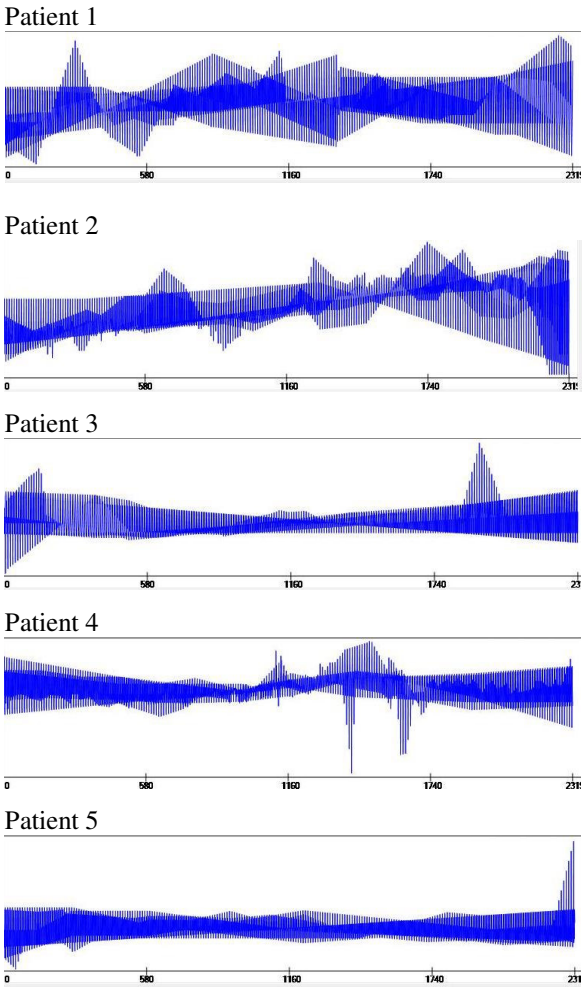


Figure 5: Multivariate representations of patients.

The MVAs clustered into two categories, but not into the desired renal and respiratory failure categories. In Figure 5, Patient 3 is the one that clustered independently of the others, but Patient 2 is the only one that experienced renal failure. Although MTSA shows promise for visualization, a different approach may be necessary for creating MTSAs on which we can relevantly categorize patients.

Related Work

Much of the research in medicine on vital signs data focuses on monitoring patients remotely and on using vital signs to determine when a patient is in duress. Duress has generally been measured by the traditional method of using preset thresholds for these values. There is also some research on personalizing medicine given a patient's genotype.

Caraça-Valente & López-Chavarrías[8] characterize possible injuries by using data mining techniques to analyze data from patients doing isokenetic exercises on a machine. They use Euclidean distance to cluster the results.

Chuah & Fu[9] use time series analysis to detect ECG anomalies. They describe a method for normalizing the time series, an adaptive window-based discord discovery (AWDD) scheme to detect abnormal heartbeats. The algorithm can be used in real time, but it is a supervised algorithm.

Seely & Macklen[10] use the science of variability analysis to describe the behavior of complex systems that have characteristic patterns of variation over time. Although their paper does not mention any data mining techniques, it does characterize the range of biological signals and describes many statistical and mathematical techniques for variability analysis.

Saeed & Mark[11] use 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, then uses a k-nearest-neighbor classification to create a predictor for hemodynamic deterioration. Their algorithm is a supervised learning algorithm.

Our research probably most closest emulates the research described by Sorani et al.[12] They 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 was available for our patients. Also, because all of the data was captured at the same frequency, no interpolation of the data was required. The patients had all experienced traumatic brain

injury (TBI) and by using multivariate time series analysis of the physiological data could be clustered into one of three categories.

Future Work

Our goal is to use larger data sets to categorize patients according to outcome. For example, patients with Traumatic Brain Injury (TBI) can be categorized into three categories: those that recover fully, those that recover with injury, and those that do not recover. We are particularly interested in patients with TBI because we expect more consistent data at more regular intervals. We aim to cluster the MTSA or SAX representations into medically relevant categories. The SAX representation's similarity to a DNA sequence suggests that we may be able to use genetic sequencing alignment methods to make the entire process unsupervised.

Since patients are in the PICU for varying lengths of time, one difficulty in comparing patients is to align these different length time series. In this paper, we used domain experts to indicate the likely start and end points that encompassed the critical "failure" period. In future work, our goal is to align the data without the use of domain experts by using sliding windows for comparison, or by using preprocessing to select likely failure regions (which should appear in the data as anomalous or transitional regions with the time series). Using data from different patient categories, we also hope to determine markers that can alert a provider to intervene in a timely fashion to change the course of a patient from heading toward a less than desirable state to desirable one.

We intend to add to our visualization the ability for the provider to set a desired state for a patient and use the visualization to compare a patient's current and past states to a desired state. To achieve this goal, we will have to make our visualization run in real time. Once the visualization is complete, we will have several anesthesiology residents handle the same simulated surgical procedure. Half will use the traditional visualization found in the OR and the others will be aided with the new MTSA visualizations and determine which group is best able to determine and correctly diagnose the critical conditions.

Conclusion

We have presented an algorithm for the creation of a Multivariate Time Series Amalgam that is comprised of interleaved univariate time series data. The visualization of the MTSA enables providers to examine a patient's overall state over time in a multivariate fashion which will improve a patient's prognosis and subsequent treatment.

References

1. Transforming Healthcare: The President's Health Information Technology Plan. http://www.whitehouse.gov/infocus/technology/economic_policy200404/chap3.html.
2. Lin J, Keogh E, Lonardi S. Visualizing and discovering non-trivial patterns in large time series databases. Information Visualization 2005; 4: 2: 61-82.
3. Keogh E, Time Series Tutorial 2006. http://www.cs.ucr.edu/~eamonn/tutorials.html ; 2006.
4. Agrawal R, Faloutsos C, and Swami A. Efficient similarity search in sequence databases. Proceedings of the Fourth International Conference on Foundations of Data Organization and Algorithms, 1993; 69-89.
5. Chan K, Fu A. Efficient time series matching by wavelet. Proceedings of the 15th International Conference on Data Engineering 1999; 126-133.
6. Morinaka Y, Yoshikawa M, Amagasa T, and Uemura S. The L-index: An indexing structure for efficient subsequence matching in time sequence databases. Proceedings of 5th Pacific-Asia Conference on Knowledge Discovery and Data 2001; 51-60.
7. Lin J, Keogh E, Wei L, Lonardi S. Experiencing SAX: a novel symbolic representation of time series. Data Min Knowl Disc 2007; 15:107-144.
8. Caraca-Valente JP, Lopez-Chiavarrias I. Discovering similar patterns in time series. Proceedings of 6th ACM SIGMOD International Conference on Knowledge Discovery and Data Mining 2000; 126-133.
9. Chuah M, Fu F. ECG Anomaly detection via time series analysis. Frontiers of High Performance Computing and Networking ISPA 2007 Workshops 2007; 123-135.
10. Seely A, Macklem P. Complex systems and the technology of variability analysis. Critical Care 2004; 8: 367-384.
11. Saeed M, Mark R. A novel method for the efficient retireval of similar multiparamenter physiologic time series using wavelet-based symbolic representations. Proceedings of AMIA Symposium 2006; 679-683.
12. Sorani MD, Hemphill III JC, Morabito D, Rosenthal G, Manley GT. New Approaches to Physiological Informatics in Neurocritical Care. Neurocritical Care 2007; 6: 1-8.