

# Similarity-Based Searching in Multi-Parameter Time Series Databases

LH Lehman, M Saeed, GB Moody, RG Mark

Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA

## Abstract

*We present a similarity-based searching and pattern matching algorithm that identifies time series data with similar temporal dynamics in large-scale, multi-parameter databases. We represent time series segments by feature vectors that reflect the dynamical patterns of single and multi-dimensional physiological time series. Features include regression slopes at varying time scales, maximum transient changes, auto-correlation coefficients of individual signals, and cross correlations among multiple signals. We model the dynamical patterns with a Gaussian mixture model (GMM) learned with the Expectation Maximization algorithm, and compute similarity between segments as Mahalanobis distances. We evaluate the use of our algorithm in three applications: search-by-example based data retrieval, event classification, and forecasting, using synthetic and real physiologic time series from a variety of sources.*

## 1. Introduction

Robust navigation and mining of physiologic time-series databases often require finding similar temporal patterns of physiological responses. In collections such as the MIMIC II database [1], we seek cases in which trends and interrelationships among vital signs exhibit patterns resembling those of a prototype (selected) case. Detection of these complex physiological patterns not only enables demarcation of important clinical events but can also elucidate hidden dynamical structures that may be suggestive of disease processes.

We describe a similarity-based searching and pattern matching algorithm that identifies time series data with similar temporal dynamics in large-scale, multi-parameter databases. We model the similarity among physiological time series with a Gaussian mixture model in which “clusters” of series with similar temporal patterns are identified. We explore the use of our pattern matching algorithm in (1) data retrieval, (2) event classification, and (3) predictive monitoring.

In the data retrieval and exploration context, we evaluate the use of our algorithm as a “search-by-example” tool.

In a collection of many time series, it seeks those that exhibit temporal patterns similar to the patterns in a given example. Such a tool may help in finding cohorts of patients with similar pathologies, and in identifying temporal patterns that may be suggestive of disease progressions.

In event classification, our goal is to differentiate among physiological trends corresponding to different clinical events, which may be useful for event detection and for alert generation for clinical decision support.

In the context of predictive monitoring, our goal is to predict significant clinical events or outcomes based on physiological measurements well before obvious signs of physiological deterioration develop in patients. The premise of our approach is that subtle patterns of vital signs and their interrelationships, common to patients with similar disease progressions, may have prognostic value.

## 2. Methods

We represent time series segments by feature vectors that reflect the dynamical patterns of single and multi-dimensional physiological time series. Features include regression slopes at varying time scales, maximum transient changes, auto-correlation coefficients of individual signals, and cross correlations among multiple signals. We model the dynamical patterns with a Gaussian mixture model (GMM) learned with the Expectation Maximization (EM) algorithm. Once the mixture model is generated, similarity between segments can be computed as Mahalanobis distances.

Mixture models are simple probabilistic models that can be used to uncover hidden structure in the data, especially in terms of unidentified subgroups. An M-component Gaussian mixture model over data  $x$  is defined as

$$P(x; \theta) = \sum_{j=1}^M P(j) N(x; \mu_j, \Sigma_j) \quad (1)$$

The parameters  $\theta$  include the prior distribution  $P(j)$ , Gaussian component means  $\mu_j$ , and covariances  $\Sigma_j$ . EM is an iterative algorithm that seeks to find the GMM parameters that maximize the log likelihood of the data [2, 3]. In the initialization step, component means are initialized to

randomly chosen data points, component covariances are set to the overall data covariance, and the prior probability for each component is  $\frac{1}{M}$ .

The E-step in each iteration  $l$  evaluates the posterior assignment probabilities  $p^l(j|i) = P(j|x_i, \theta^l)$  based on the current setting of the parameters  $\theta^l$ , where  $j = 1, \dots, M$ , and  $i = 1, \dots, N$ . For each data point, we compute the posterior probability that it is generated from each of the components. The M-step updates the parameters  $P(j)$ ,  $\mu_j$ , and  $\Sigma_j$  based on the posterior probabilities  $p^l(j|i)$  from the E-step. To overcome numerical problems, we use a regularized EM algorithm [3, 4].

The algorithm stops when the relative change in log-likelihood falls below some small amount  $\epsilon$  or when a specified maximum number of iterations is reached. Since EM can get stuck in locally optimal solutions, we re-run EM multiple (typically 100 to 1000) times, each with a different initialization of the component means. The one with the maximum log-likelihood of the fitted mixture of Gaussians is returned by the algorithm as the final mixture model for a given  $M$ .

We use a K-nearest neighbor based classification rule, in which the class/type of a temporal pattern in question is determined based on its similarity to libraries of patterns associated with known events. Given a segment  $i$ , we retrieve its  $K$  nearest neighbors (in terms of Mahalanobis distances) from its assigned Gaussian component. A threshold-based rule is used for binary classification as in [5]. We use a majority based rule for multi-class classification. Upon a tie,  $i$  is classified using the “closest” neighbor from the competing classes.

### 3. Evaluation

We tested our algorithm’s performance in retrieval, classification, and prediction accuracy using three data sets. First, we utilized a synthetic benchmark data set consisting of 6 classes with 100 representative examples from each class. In another experiment, we examined heart rate and blood pressure time series from tilt tests of ten healthy human subjects (10 hours in all, containing 120 responses to six different interventions). Finally, we assessed our algorithm in predicting recurrent hypotensive episodes (after the withdrawal of vasoactive medications) using the MIMIC II database.

#### 3.1. Search by example

We applied our algorithm to the task of finding time series in the Synthetic Control Chart Time Series data set from UC-Irvine [6] that were similar to a randomly chosen example from that set. The data set consists of 600 artificially generated time series, forming 6 classes (100 examples of each class) of similar time series: normal,

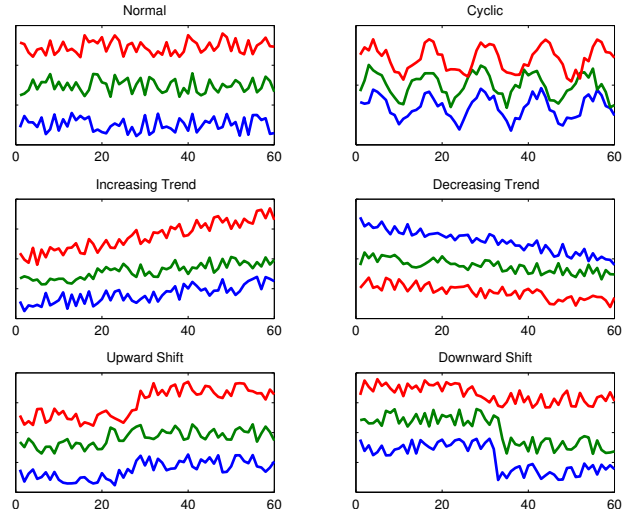


Figure 1. Example patterns from the synthetic data set. Three examples from each of the six classes are shown.

cyclic, increasing trend, decreasing trend, upward shift, and downward shift (see figure 1).

By using each of the 600 synthetic time series in turn as an example to be matched, our algorithm found the 99 best matches to the example from the remaining 599 series. Using a six-component GMM to find the best K-nearest neighbor matches for a randomly selected case, the retrieval accuracy was 95% when K is 99 (in other words, about 95 of the 99 nearest neighbors, on average, belonged to the same class as the example).

#### 3.2. Classification

Using the same data set and six-component GMM as in the search by example application, we evaluated the algorithm’s classification performance. With one-nearest neighbor (1-NN) as a classification rule, the algorithm was able to classify 588 out of 600 synthetic time series correctly (98% classification accuracy).

We conducted another study of classification using a real data set that included ten hours of hemodynamic (heart rate and blood pressure) measurements from ten healthy human subjects undergoing 6 different types of tilt table interventions [7]: slow tilting up/down (75 degrees in 50 seconds), rapid tilting up/down (75 degrees in 2 seconds), and transitions between standing and supine positions (transitions in less than 2 seconds). Interventions were 5 minutes apart. Each subject underwent 12 interventions (2 per type) over a one-hour period.

In this second study, our algorithm classified the transient physiological responses of the subjects that corresponded to the six interventions, with the goal of being able to identify which intervention evoked a given re-

Table 1. Recall and precision using 1-nearest neighbor classification with a 6-component GMM. Overall classification accuracy is 0.88.

	Tilt Up	Tilt Down	Stand Up	Lie Down
Recall	0.88	0.88	0.85	0.90
Prec.	0.92	0.85	1.00	0.75

sponse.

Classification performance is measured in terms of class recall and precision. Recall and precision for class  $i$  are measured as  $\frac{TP_i}{(TP_i+FN_i)}$ , and  $\frac{TP_i}{(TP_i+FP_i)}$  respectively, where  $TP_i$ ,  $FN_i$  and  $FP_i$  are the number of true positives, false negatives, and false positives in class  $i$  respectively. The overall classification accuracy is defined as the fraction of events that the algorithm classifies correctly.

We extracted 120 three-minute segments of heart rate and mean arterial blood pressure (each sampled at 2 Hz) from the ten hours of recordings. Each segment started one minute before the intervention. We constructed feature vectors from the HR and MAP signals using regression slopes at varying time scales (ranging from 10 seconds to 3 minutes), maximum transient increase and decrease in a 5 second interval, and auto-correlation with a 5-second lag. K-nearest neighbor classification based on majority rule was used.

We observed that the responses were affected only in subtle ways by the speed of the tilt, so our algorithm initially grouped the responses into only four classes (tilt up, tilt down, stand up, lie down). Table 1 reports the recall and precision for each class obtained using a 6-component GMM with a 1-NN rule. The overall classification accuracy is 0.88. In the more challenging task of six-way classification (differentiating between slow and fast tilts as well) using the same 6-component GMM, we obtained classification accuracies of 0.67 with a 1-NN rule, and 0.71 with an 8-NN rule. As part of our future work, we are exploring feature sets that can be used to differentiate a wider variety of temporal patterns.

### 3.3. Forecasting

Finally, we applied our algorithm to the task of forecasting hypotensive episodes in intensive care unit patients. Early warning of these life-threatening events can allow the medical staff to take precautionary measures. Using cases from the MIMIC II database in which recurrent hypotensive episodes were observed, we attempted to identify features of the physiologic time series that can be used to distinguish between patients who stabilized from a previous hypotensive episode and those who deteriorated fur-

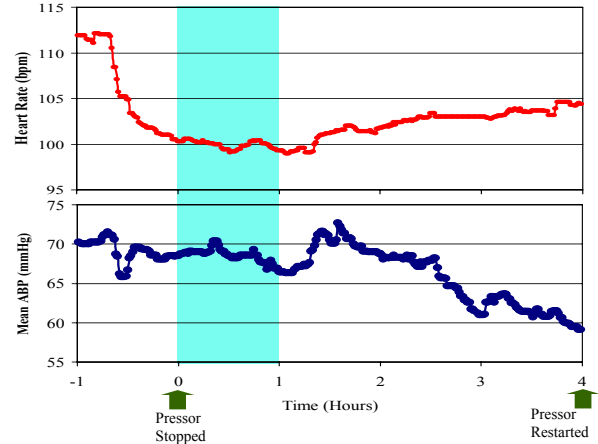


Figure 2. An excerpt of a MIMIC II record from a patient with recurrent hypotension after pressor withdrawal. Pressor was withdrawn at time 0. The patient required re-administration of pressor therapy approximately 4 hours later. Our current study used the HR and MAP measurements during the one-hour period (shaded area) immediately following the pressor withdrawal to predict whether the patient would become hypotensive again.

ther to develop recurrent hypotensive episodes.

In this pilot study, we focused on patients who received intravenous vasoactive medication (pressor) therapies. Our goal was to utilize the physiological measurements obtained during the first hour of pressor withdrawal to predict which patients would successfully wean off of pressors, and which ones would develop recurrent hypotension within 2 to 6 hours after withdrawal from their previous pressor administration. Figure 2 shows an excerpt of a MIMIC II record of a patient with recurrent hypotension after pressor withdrawal. The shaded area represents the data segment used for the prediction task.

Two cohorts of patients were selected from the MIMIC II database: the *stable* group (118 instances of pressor therapies, 118 unique patients) consisted of patients who were successfully weaned from vasoactive medications without the need for successive pressor therapies. The *unstable* group (109 instances of pressor therapies, 85 unique patients) consisted of patients who required re-administration of intravenous pressor therapies 2 to 6 hours after the withdrawal of their previous pressor treatment.

We preprocessed the time series of heart rate (HR) and mean arterial blood pressure (MAP) measured at one-minute intervals with median filters and removed noisy samples that fell beyond physiological bounds. Linear interpolation was used to fill in the missing values. Feature vectors were constructed from one-hour segments of HR and MAP time series beginning immediately after the withdrawal of the pressor therapy. Features extracted in-

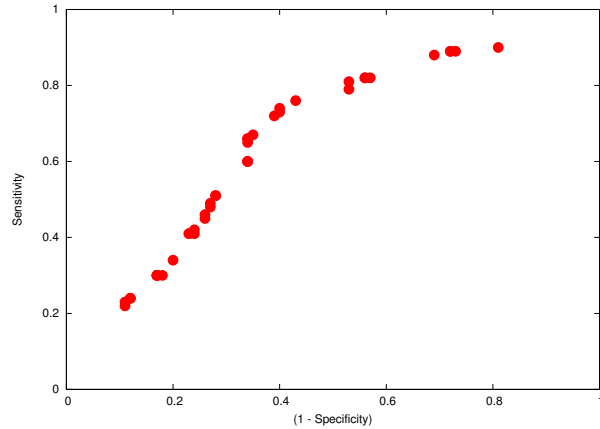


Figure 3. ROC for predicting outcome between *stable* and *unstable* groups, using a 26-component GMM with a 23-NN classification rule, generated by varying the unstable threshold ( $0.3 \leq \rho \leq 0.7$ ).

cluded 15- and 60-minute regression slopes, maximum transient increase and decrease, auto-correlation, and cross correlation between MAP and HR.

We used a threshold-based K-nearest neighbor classification rule as in [5]. The classification was based on an unstable threshold setting  $\rho$ , where ( $0 < \rho < 1$ ). A patient was classified as *unstable* if at least  $\rho K$  of the  $K$  nearest neighbors were *unstable*; otherwise, the patient was classified as *stable*.

Sensitivity ( $\frac{TP}{TP+FN}$ ) and specificity ( $\frac{TN}{TN+FP}$ ) were used to measure prediction accuracy, where TP (true positive) and FP (false positive) were the numbers of correctly and incorrectly labeled unstable events, and TN (true negative) and FN (false negative) were the numbers of correctly and incorrectly labeled stable events.

Ten-fold cross validation was used for evaluation. The mixture model was constructed using the training set, and the model with the best average test error was reported. Figure 3 shows the Receiver Operating Curve (ROC) for forecasting using a 26-component mixture model. The area under the ROC is 0.67. A sensitivity of 0.74, with a specificity of 0.60, was obtained at  $K = 23$ , and unstable threshold  $\rho = 0.41$ . As part of our future work, we plan to use physiological measurements obtained *before* the withdrawal of the pressors to predict recurrent hypotensive episodes.

#### 4. Conclusions and future work

We developed a similarity-based searching and pattern matching algorithm and evaluated its use in searching by example, event classification, and forecasting. Using a synthetically generated time-series data set, we demonstrated that our algorithm achieves high accuracy in search-

by-example retrieval and in event classification. We also demonstrated the algorithm's potential use in event classification and forecasting using real physiological measurements from tilt tests and the MIMIC II database. The performance analysis suggests that the robustness of forecasting algorithms is highly dependent on the feature vectors that are used for training various machine learning algorithms. Future work will focus on developing compact feature sets that best characterize the salient physiologic dynamics that distinguish different physiologic states in ICU patients.

#### Acknowledgements

The authors would like to acknowledge Dr. Thomas Heldt for providing the tilt data set, and for his valuable input to the project. We also thank Prof. Tommi Jaakkola for valuable discussions and the MIT course 6.867 teaching staff for sharing code for the EM algorithm. This work was funded in part by the National Institute of Biomedical Imaging and Bioengineering and by the National Institute of General Medical Sciences, under NIH cooperative agreement U01-EB-008577 and NIH grant R01-EB001659.

#### References

- [1] Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Computers in Cardiology 2002*; 29:641–644.
- [2] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B June 1977*;39:1–38.
- [3] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
- [4] Jaakkola T. Course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (<http://ocw.mit.edu>), Massachusetts Institute of Technology.
- [5] Saeed M. *Temporal Pattern Recognition in Multiparameter ICU Data*. Doctoral dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, June 2007.
- [6] Hettich S, Bay SD. The UCI KDD archive: Synthetic control chart time series. <http://kdd.ics.uci.edu>, 1999. Irvine, CA: University of California, Department of Information and Computer Science.
- [7] Heldt T. *Computational Models of Cardiovascular Function During Orthostatic Stress*. Doctoral dissertation, Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA, September 2004.

Address for correspondence:

Li-wei H. Lehman  
 Harvard-MIT Health Sciences and Technology  
 E25-505, Cambridge, MA 02139 USA  
 lilehman@mit.edu