Unsupervised Pattern Discovery in Electronic Health Care Data Using Probabilistic Clustering Models

Benjamin M. Marlin
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA, USA
marlin@cs.umass.edu

David C. Kale
Whittier Virtual Pediatric Intensive Care Unit
Children's Hospital Los Angeles
Los Angeles, CA, USA
dkale@chla.usc.edu

Robinder G. Khemani
Department of Anesthesiology and Critical Care Medicine, Children's Hospital Los Angeles
Los Angeles, CA, USA
rkhemani@chla.usc.edu

Randall C. Wetzel
Department of Anesthesiology and Critical Care Medicine, Children's Hospital Los Angeles
Los Angeles, CA, USA
rwetzel@chla.usc.edu

ABSTRACT

Bedside clinicians routinely identify temporal patterns in physiologic data in the process of choosing and administering treatments intended to alter the course of critical illness for individual patients. Our primary interest is the study of unsupervised learning techniques for automatically uncovering such patterns from the physiologic time series data contained in electronic health care records. This data is sparse, high-dimensional and often both uncertain and incomplete. In this paper, we develop and study a probabilistic clustering model designed to mitigate the effects of temporal sparsity inherent in electronic health care records data. We evaluate the model qualitatively by visualizing the learned cluster parameters and quantitatively in terms of its ability to predict mortality outcomes associated with patient episodes. Our results indicate that the model can discover distinct, recognizable physiologic patterns with prognostic significance.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Algorithms, Experimentation

Keywords

Probabilistic Models, Clustering, Time Series, Electronic Health Records, Critical Care, Pediatrics

1. INTRODUCTION

Providing critical care requires highly time-sensitive decision making. Patients in the *intensive care unit* (ICU) produce a stream of data from a variety of physiologic monitors, laboratory tests and subjective assessments. Small changes in the data over time can precede large catastrophic events. Recognizing potentially modifiable patterns in physiology can lead to better treatment and more efficient care, and potentially prevent death and disability. Intensive care physicians are charged with integrating and interpreting this stream of data in the context of their training and years of experience. This process may be understood as recognizing similarities between new patients and representative previous cases. A long standing goal in the area of medical informatics is the development of decision support tools that can automate aspects of this process, but efforts have largely focused on rules-based expert systems, which are ill-suited to the pattern recognition nature of the task.

More modern approaches based on techniques from computational statistics and machine learning hold great promise but require the availability of large amounts of clinical data for estimating models [25]. There are currently a variety of efforts underway to make massive stores of high dimensional, granular clinical data easier to collect, manage and share in clinical and research settings [19, 2, 6]. Until recently, such repositories were highly uncommon, since building them required either prospective collection or labor-intensive review of paper charts. With the adoption of electronic health care records (EHRs), much of this data is already captured and available digitally, increasing the feasibility of developing rich data repositories. Nevertheless, the cost and risk of building, maintaining, and sharing these repositories are high enough that many institutions are reluctant to invest in them. If researchers can demonstrate the potential of analyzing such large stores of clinical data, regulators and institutions may be motivated to adopt policies that encourage the development and sharing of such repositories.

However, analyzing real clinical data presents challenges that go far beyond data acquisition and management. First, all data must be treated as fundamentally uncertain. Many observations are recorded manually, introducing the possibility of various kinds of human error. Even physiological measurements that can be captured automatically from monitoring equipment are subject to uncertainty introduced by sensor noise, malfunctions and other unexpected events.

IHI'12, January 28–30, 2012, Miami, Florida, USA.

Second, data are recorded at a wide variety of sampling frequencies, ranging from high frequency heart rate waveforms to the results of invasive lab tests ordered by clinical staff.

In this paper, we analyze moderate-frequency EHR data collected over a ten year period in the pediatric intensive care unit (PICU) at Children's Hospital Los Angeles. This data is sparse, high-dimensional and uncertain, necessitating the use of specialized algorithms derived from probabilistic models. We describe a probabilistic clustering model that includes an empirical prior distribution on the model parameters designed specifically to mitigate the sparsity inherent in physiological data extracted from real EHRs. Using exploratory analysis and visualization, we show that the clusters are associated with recognizable physiologic and diagnostic patterns. Finally, we demonstrate that the clusters have prognostic significance by constructing cluster-based mortality prediction models that achieve superior performance compared to treating all patient episodes as a single group.

2. RELATED WORK

There is a tremendous body of work in pediatric critical care focused on using physiologic measurements, such as heart rate and blood pressure, to characterize critical illness severity. Much of this work has focused on the construction, validation and application of *severity of illness* (SOI) scores, such as the Paediatric Index of Mortality 2 (PIM) [22], the Paediatric Logistic Organ Dysfunction (PELOD) score [10] [9] and the Pediatric Risk of Mortality III (PRISM III) [17] score. Their development preceded the common use of digital data in ICUs and was largely driven by two priorities: early detection of severe illness (within hours of admission to a unit) and economy of data required for calculation. These scores are routinely used to stratify patient risk to benchmark PICU performance. They are not designed, however, for making predictions about or informing the care of individual patients [14]. Moreover, they do not take advantage of the abundant, granular data found in EHRs. From a machine learning perspective, these scores can be viewed as an application of supervised classification using features carefully chosen by experts. They largely ignore the temporal nature of the data and are sensitive to missing values.

More recent work attempts to take advantage of the rich data captured by medical devices and EHRs, particularly high frequency time series recorded from bedside monitors. Lehman, et al., combine manually designed features meant to capture temporal information (e.g., gradients and trends) with a Gaussian mixture model for clustering episodes from the MIMIC-II database [8], showing great success in "search by example" and classification tasks. The main drawback of this work is the reliance on manually defined features.

Other recent work has focused on the problem of unsupervised discovery of meaningful patterns and features from raw data. One example is Lin and Li, who convert continuous values to discrete symbols and then build "bag-of-words" representations of each physiologic time series, similar to the way documents are modeled in information retrieval tasks [11]. They demonstrate that this representation outperforms other published results in classification and clustering tasks performed on data sets from Physionet. While this approach captures certain higher-level structure of physiological time series, it discards important information related to temporal order. Saria, et al., propose another method that directly models the heterogeneous, temporal nature of physiological time series using a switching latent "topic" model similar to those used widely in natural language processing [20]. They show that such a model can be used to construct interesting features for use in other tasks, such as their proposed neonatal SOI score, PhysioScore [21].

This more recent work makes effective use of the rich data generated in ICUs, but it depends on the availability of complete, high frequency time series. While stores of such data are increasingly common at large, cutting edge research institutions, it is still largely uncaptured and unavailable at many hospitals and clinics, even those who have adopted EHR systems. By contrast, our work specifically addresses the uncertain and sparsely sampled time series data that is common in EHRs, as well as the variety of challenges this data presents.

3. DATA SET

Recording clinical data during the delivery of care (not prospectively or intentionally for research) has several key implications that make working with it challenging. First, it is *incomplete*. We do not have observations of every variable for every patient. Rather, we have observations of only those phenomena recorded by caregivers during treatment. For example, if a clinician chooses not to order blood tests, we may not have measurements of pH or glucose. This constitutes a potential source of non-random missing data. Next, the data are sampled in a *sparse, non-uniform* manner. The rate at which observations are recorded varies depending on the variable, the way measurements are made, which caregiver is recording measurements and possibly the patient's level of illness (for example, heart rate may recorded more frequently when a patient's condition becomes more critical). This constitutes a potential source of *sample selection bias* [5, 23]. Finally, the data are subject to various forms of *uncertainty*, including errors introduced during manual recording by clinical staff and technicians, as well as measurement noise and failures of automated monitoring devices. Nevertheless, this kind of observational data is increasingly abundant and is highly typical of the information being collected in most production EHR systems today. The ability to estimate accurate models from such data is of enormous value.

In this work, we use a novel data set collected from the PICU EHR archive at the Children's Hospital of Los Angeles. This data set contains over 10,598 PICU patient episodes collected over a ten year period and includes essentially all PICU episodes that could be reliably extracted and verified from the available EHRs. Though the data set includes demographics, outcomes, diagnostic codes, and other annotations, we focus on learning clustering models from the physiological time series data only, including timestamped measurements of thirteen different variables. These variables were chosen for availability and ease of extraction and because they have known prognostic value, but the model generalizes to any set of available physiologic variables. The physiological data exhibit all of the problems described above (incompleteness, sampling irregularity, uncertainty), but are also extremely rich. A summary of the thirteen variables included in the data set can be found in Table 1. Column one refers to the abbreviation for the variable, column two gives the full variable name and column three is the average number of measurements of that variable per day.

Abbrev.	Description	Msmts per day
SpO2	Pulse oximetric saturation	31.04
HR	Heart rate	29.78
RR	Respiratory Rate	29.70
sBP	Systolic blood pressure	23.52
dBP	Diastolic blood pressure	23.50
EtCO2	End-tidal carbon dioxide	13.85
Temp	Temperature	11.68
TGCS	Total Glascow coma score	11.24
CRR	Peripheral capillary refill rate	11.18
UO	Urine output	9.50
FiO2	Fraction inspired oxygen	5.17
Gluc	Glucose	2.06
pH	pH	1.50

Table 1: Abbreviations, descriptions and measurement frequencies for each of the 13 variables included in the data set.

3.1 Preprocessing

The temporal nature of this data, particularly with respect to non-uniform sampling and large variance in episode length, poses significant challenges to traditional statistical techniques. We select a subset of the data and apply preprocessing steps to simplify the subsequent analysis of the data with the understanding that these choices discard potentially useful information.

First, we automatically discard anomalous measurements with values outside a valid range defined by clinical experts. While these measurements may contain some information, more sophisticated models would be required to separate it out from error and noise.

Second, we choose to focus on the first 24 hours of measurements only, ignoring any additional measurements taken afterward. While this discards potentially useful information from later in the episode, it also reduces the dimensionality of our time series and enables the application of simpler modeling techniques. Further, models based on a limited time horizon have enormous practical importance, since the ability to predict patient outcomes based on a minimal number of measurements is a key problem.

Third, we choose to discretize time into hour-long intervals. The value of a variable for each interval is the mean of all measurements taken during that hour. This is a common approach to time series dimensionality reduction, described as *Piecewise Aggregate Approximation* (PAA) by Keogh, et al. [7]. This simplifies our analysis by enabling the use of models that operate on fixed-dimensional vectors, but it also clearly discards potentially useful information and introduces additional challenges. Foremost among these is the introduction of *missing data* in the form of intervals with no measurements. There are a variety of strategies for handling such missing data including case-deletion and imputation. However, we choose to make an assumption that the missing data is *missing at random* (MAR) and apply probabilistic models that can efficiently deal with missing data under this assumption. It is important to state here that this assumption does not represent an actual belief about the nature of the data, but simply reflects a choice about the family of models we consider in this work.

It is also worth noting that our final data set is fully de-identified, as our policy is to be especially conservative about patient privacy and data security. All protected health information (PHI) is stripped and potential "outlier" episodes (e.g., patients with especially long PICU stays or unusual demographics or diagnoses) are removed entirely. We also convert all dates and times to timestamps *relative to time of admission* (e.g., birthday converted to age in months at time of admission, measurement timestamps converted to "minutes since time of admission", etc.). We adopted this policy, with full approval from the CHLA *Institutional Review Board* (IRB), in order to fully protect patient privacy while still enabling our own research, as well as the sharing of data with other researchers in the future.

4. PROBABILISTIC CLUSTERING

The final preprocessed data set described in Section 3 is moderately high-dimensional and contains observations that are incomplete and potentially uncertain. Well-known clustering methods like K-means clustering [13] and hierarchical clustering [24] rely on the availability of a distance metric defined on the space of the data. Unfortunately, these metrics can not usually deal with missing measurements in the data vectors. By contrast, probabilistic clustering models, also referred to as mixture models [15], can deal with missing data very efficiently under certain assumptions [4]. In this section, we first describe the basic model we use in this work: a mixture of diagonal covariance Gaussians. We next describe how to extend the basic model using an informative prior distribution designed specifically to make the model more robust to sparsely-sampled data. Finally, we present an algorithm for estimating the model and report the results of estimating the model on the CHLA data set.

4.1 Notation

In describing the model, we will let N indicate the number of data cases, each of which corresponds to a single patient episode. We let V indicate the number of physiological variables and T indicate the number of measurement time points. The CHLA data set used in this paper has $V = 13$ physiological variables and $T = 24$ measurement time points, one per hour for 24 hours.

We denote the data matrix by $\boldsymbol{X}$ and the entry for data case n, variable v and time point t by X_{nvt}. In a probabilistic model, each measurement X_{nvt} is considered to be a random variable. An instantiation or value of the random variable X_{nvt} is denoted by x_{nvt} (in general, we will use capital letters to indicate random variables and lower case letters to indicate instantiations or values of random variables). We use the notation $\boldsymbol{X}_{nv}$ to indicate the vector-valued random variable corresponding to all T measurements of variables v for data case n, and $\boldsymbol{x}_{nv}$ to indicate an instantiation of this random variable. When the data matrix contains missing values, it is useful to consider a companion matrix of binary response indicator variables $\boldsymbol{R}$ of the same size as $\boldsymbol{X}$. We set $r_{nvt} = 1$ if x_{nvt} is observed and $r_{nvt} = 0$ if x_{nvt} is not observed.

4.2 Model Description

The diagonal covariance Gaussian mixture model is a probabilistic clustering model for real-valued data. It assumes that a fixed, finite number of mixture components or clusters K underlie the data. The model can be thought of as a stochastic process for generating completely observed

data cases. The generative process for data case n begins by selecting a cluster from a prior distribution over clusters $P(Z_n = k) = \theta_k$. The random variable Z_n is referred to as a latent or hidden variable and indicates which cluster data case n belongs to. The distribution over clusters is simply a discrete distribution parameterized by θ_k. Given the sampled value $Z_n = k$, a value for x_{nvt} is sampled independently for each measurement variable X_{nvt} from a univariate Gaussian (normal) distribution $\mathcal{N}(\mu_{kvt}, \sigma_{kv}^2)$ associated with cluster k. The mean of this Gaussian distribution is μ_{kvt} while σ_{kv} is the standard deviation. Note that we assume the cluster mean varies as a function of time, while the cluster standard deviation is constant through time. We summarize the basic model below.

$$P(Z_n = k|\theta) = \theta_k \tag{4.1}$$

$$P(X_{nvt} = x_{nvt}|\mu_{kvt}, \sigma_{kv}) = \mathcal{N}(x_{nvt}; \mu_{kvt}, \sigma_{kv}^2) \tag{4.2}$$

The probability of a data case under this model is a mixture over the K clusters. When a data case is incompletely observed, the missing data can be analytically marginalized away under the assumption that the missing data is *missing at random* (MAR) [12]. There is no need to perform any explicit imputation of missing data values. We give the probability of an incompletely observed data case under the MAR assumption below. The effect of r_{nvt} in the exponent is to include a contribution from x_{nvt} only if x_{nvt} is observed.

$$P(\boldsymbol{x}_n|\boldsymbol{r}_n, \theta, \mu, \sigma) = \sum_{k=1}^{K} \theta_k \prod_{v=1}^{V} \prod_{t=1}^{T} \mathcal{N}(x_{nvt}; \mu_{kvt}, \sigma_{kv}^2)^{r_{nvt}} \tag{4.3}$$

Given an incompletely observed data case $(\boldsymbol{x}_n, \boldsymbol{r}_n)$, we will need to infer the posterior distribution over the latent variables Z_n. We can use this posterior distribution to make a hard assignment of data cases to clusters after the model is learned if needed. This posterior distribution $q_{nk} = P(Z_n = k|\boldsymbol{x}_n, \boldsymbol{r}_n, \theta, \mu, \sigma)$ can be found using Bayes rule and the definition of the model as follows:

$$q_{nk} = \frac{\theta_k \prod_{v=1}^{V} \prod_{t=1}^{T} \mathcal{N}(x_{nvt}; \mu_{kvt}, \sigma_{kv}^2)^{r_{nvt}}}{\sum_{k=1}^{K} \theta_k \prod_{v=1}^{V} \prod_{t=1}^{T} \mathcal{N}(x_{nvt}; \mu_{kvt}, \sigma_{kv}^2)^{r_{nvt}}} \tag{4.4}$$

4.3 Model Estimation

Maximum likelihood (ML) estimation of the parameters in a Gaussian mixture model can be performed using an Expectation Maximization (EM) algorithm [3]. The EM algorithm for a basic Gaussian mixture model with incomplete data has previously been described by Ghahramani and Jordan [4]. This algorithm alternates between computing the posterior distribution over the clusters for each data case (E-Step) and updating the parameters of the cluster distributions (M-Step).

However, in the current setting where we have a large amount of missing data, standard ML learning can behave quite poorly. When the data is highly sparse, the mean values for individual time points in small clusters may be estimated based on very few observations, making them very noisy. The situation with respect to the standard deviation parameters is less severe due to the fact that we tie these parameters across time. However, if a particular variable has very few observations over all time points in a particular cluster, the standard deviation estimate can also behave poorly. One solution to these problem is to include an informative prior distribution on the model parameters and estimate the model using maximum a posteriori (MAP) estimation.

Algorithm 1 MAP EM Algorithm for Diagonal Covariance Gaussian Mixture with Empirical Smoothing Prior

for $i = 1$ to I **do**
  for $n = 1$ to N, $k = 1$ to K **do**
    $q_{nk} \leftarrow P(Z_n = k|\boldsymbol{x}_n, \boldsymbol{r}_n, \theta, \mu, \sigma)$
  end for
  for $k = 1$ to K, $v = 1$ to V **do**

$$\theta_k \leftarrow \frac{1}{N} \sum_{n=1}^{N} q_{nk}$$

$$\sigma_{kv}^2 \leftarrow \frac{N_0 \sigma_{0v}^2 + \sum_{n=1}^{N} \sum_{t=1}^{T} r_{nvt} q_{nk} (x_{nvt} - \mu_{kvt})^2}{N_0 + \sum_{n=1}^{N} \sum_{t=1}^{T} r_{nvt} q_{nk}}$$

$$\mu_{kv} \leftarrow (\Sigma_{0v}^{-1} + \sigma_{kv}^{-2} \sum_{n=1}^{N} q_{nk} \mathrm{diag}(\boldsymbol{r}_{nv}))^{-1} \cdot (\Sigma_{0v}^{-1} \mu_{0v} + \sigma_{kv}^{-2} \sum_{n=1}^{N} q_{nk} \mathrm{diag}(\boldsymbol{r}_{nv}) \boldsymbol{x}_n)$$

  end for
end for

Since the mean parameters $\mu_{kv} = [\mu_{kv1}, ..., \mu_{kvT}]$ themselves represent a time series, a natural assumption is that the mean parameters should exhibit a degree of smoothness with respect to time. Further, when a mixture component contains few observations, we would like its parameters to fall back to the overall mean of the data. It is possible to achieve these two effects simultaneously using a kernel-based Gaussian prior on the mean parameters that enforces smoothness with respect to time. To define this prior, we take an "empirical Bayes" approach. We begin by computing the empirical means μ_{0v} and standard deviations σ_{0v} from all of the available data. We then define a similarity kernel $\mathcal{K}(t, t')$ between time points t and t'. In this work, we use a square-exponential kernel $\mathcal{K}$ with parameters a_0 and b_0 as seen below. We use this kernel matrix to define the prior covariance matrix Σ_{0v}.

$$\mathcal{K}_{tt'} = b_0 \exp(-a_0 (t - t')^2) \tag{4.5}$$

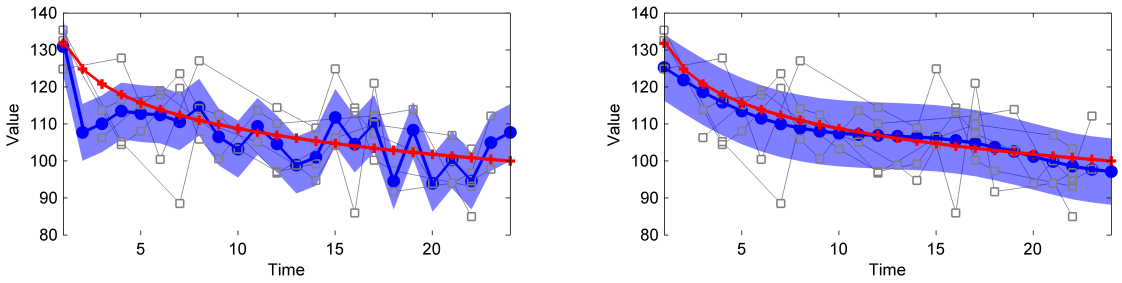
$$\Sigma_{0v} = \sigma_{0v} \mathcal{K}_{tt'} \tag{4.6}$$

The prior distribution on the cluster means is then simply a Gaussian distribution with mean μ_{0v} and covariance matrix Σ_{0v}. The construction of this Gaussian prior distribution over the mean parameter vectors is closely related to the construction of Gaussian processes, which yield a similar prior distribution over functions (as opposed to vectors) [18].

$$P(\mu_{kv}|\mu_{0v}, \Sigma_{0v}) = \mathcal{N}(\mu_{kv}; \mu_{0v}, \Sigma_{0v}) \tag{4.7}$$

We also construct an empirical prior distribution on the standard deviations parameters. We parameterize the prior in terms of the empirical variance σ_{0v} and an equivalent sample size N_0. The form of this prior corresponds to an inverse Gamma distribution on the variance parameters.

$$P(\sigma_{kv}|N_0, \sigma_{0v}) \propto \frac{1}{\sigma_{kv}^{N_0}} \exp(-\frac{N_0 \sigma_{0v}}{2\sigma_{kv}^2}) \tag{4.8}$$



(a) ML Estimate (b) MAP Estimate

Figure 1: These plots show the maximum likelihood and maximum a posteriori estimates of the mean and standard deviation parameters estimated on a small synthetic data set. The true mean function is shown in red. The estimated mean function is shown in blue. The shaded region indicates a one-standard-deviation interval about the estimated mean function. The individual data cases are shown as thin lines with square markers indicating the observed measurement points.

We give a MAP EM algorithm for the diagonal Gaussian mixture model augmented with this empirical prior in Algorithm 1. The MAP E-Step is identical to ML E-Step for a standard diagonal Gaussian mixture model. The MAP M-Step updates differ from the ML M-Step updates since they take the prior on the parameters into account, leading to regularized parameter estimates. In particular, we note that in the absence of any data for variable v in cluster k, the estimated mean parameters will revert to the overall mean μ_{0v} for variable v. Similarly, the estimated standard deviation will revert to the overall standard deviation σ_{0v} for variable v in the absence of any data about variable v. As the amount of data available to estimate any given parameter increases, the estimate of that parameter will approach its ML estimate.

Figure 1 illustrates the importance of the smoothing prior when estimating the mean parameters from sparse time series data. We construct a small synthetic data set using the true mean function shown as the red curve (+ markers) in each plot. We sample five data cases from a Gaussian distribution with this mean function with unit variance. We randomly sample the response indicators for each data case with probability 0.5. We set the measurement at each time point to be missing or observed according to the sampled response indicator vectors. The individual data cases are shown as square markers connected by dashed lines in each plot. We consider estimating the mean and standard deviation parameters from this data using both maximum likelihood estimation (neglecting the influence of the prior distribution) and maximum a posteriori estimation (taking the prior into account). We estimate the model using a single cluster. The estimated mean function is shown as a blue line (circular markers). The shaded region indicates a one-standard deviation interval about the mean. We can easily see that the MAP estimate under an appropriate choice of prior has the desired effect of smoothing the estimated means while the maximum likelihood estimate is very noisy due to the small number of observations at most time points.

The MAP EM algorithm is typically run for a fixed number of iterations I or until convergence of the probability of the data under the model. 25 to 50 iterations are usually sufficient for the algorithm to converge. In the experiments that follow, we used an initial cross validation search with the five cluster model to find hyper-parameter values that result in good predictive performance. The final selected values used for all models were $a_0 = 0.002$, $b_0 = 0.1$, $N_0 = 0.001$.

4.4 Clustering Experiments

We apply the model to cluster the pediatric intensive care patient episodes contained in the CHLA data set. To evaluate the model, we adopt a five-fold cross validation procedure. We partition the available episodes into five equal-sized blocks at random. We use one block for testing, one block for validation and the three remaining blocks for training. We rotate the blocks used for testing and validation, resulting in five different train/test/validation splits. We estimate clustering models with 5, 10 and 20 clusters.

Figure 2 shows statistics for each of the models learned on the first training split. The first row shows the size of each cluster in each model. We also use training-set mortality outcomes to estimate a probability of mortality for each cluster and each model, as shown on the second row of Figure 2. Recall that the cluster model estimation and inference algorithms are both based on raw physiological measurements and do not have access to mortality outcomes. We see that the probability of mortality varies widely between clusters, indicating that the underlying model has the ability to stratify training episodes by mortality. Note that the overall mortality probability is indicated by the solid horizontal line in each figure. We see that some clusters have highly elevated mortality probability relative to the baseline (although these clusters are inevitably small in size), while others have significantly decreased mortality probability.

In Figure 3, we visualize the mean and standard deviation parameters for the clusters with the lowest and highest mortality probabilities found using the 20 cluster model on the first split of training data. For each variable, the blue line shows the mean parameters for that variable, while the standard deviation parameters are represented by the filled blue region. The region extends above and below the mean parameter at time t by the estimated standard deviation. We refer to this visualization as a *physiome*. We can immediately see that the high-mortality cluster is associated with a depressed TGCS score and elevated heart rate, while the low-mortality cluster exhibits the exact opposite pattern. Other clusters within this model (as well as models

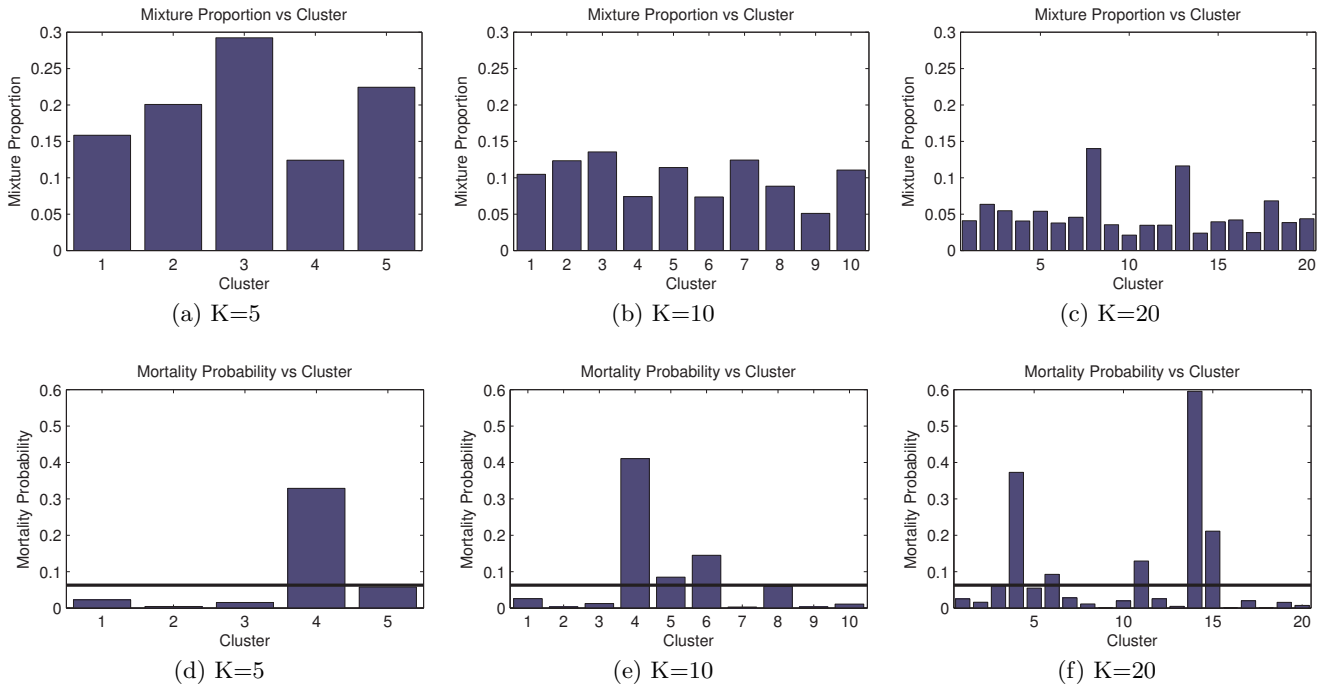


Figure 2: The first row shows the proportion of training cases assigned to each cluster for models with $K = 5, 10, 20$ clusters. The second row shows the mortality probability associated with each cluster for models with $K = 5, 10, 20$ clusters. The solid line in each figure in the second row shows the overall mortality probability.

using different numbers of clusters) similarly exhibit differentiated physiologic patterns.

5. MORTALITY PREDICTION

The cluster analysis presented in the previous section indicates that the clusters learned from the CHLA data exhibit recognizable physiological patterns and correlate well with mortality outcomes. To formalize the extent to which the inferred clusters exhibit prognostic significance, we consider using them to predict mortality outcomes for individual patient episodes.

We consider two baseline mortality prediction models. The first is a simple Bernoulli model that predicts mortality with a constant probability π. The second model is a standard linear logistic regression model. The logistic regression model requires that fully observed feature vectors be available for training the model. Inspired by feature extraction often employed in developing SOI scores, we extract the highest and lowest measurement from the first 24 hours of data for each of the 13 physiological variables, resulting in a 26-dimensional feature vector. We also include the patient's age as an additional feature variable.

We consider extending both of these baseline mortality prediction models to the clustering case by estimating them on a per-cluster basis. We use the fact that we have a probabilistic clustering model to assign each training case to each cluster according to the posterior probability $P(Z = k|\boldsymbol{x}, \boldsymbol{r})$. We then estimate one mortality prediction model for each cluster using the posterior weighted data. To deal with missing data, we use the corresponding clustering model to impute all missing observations by replacing them with their expected values under the model. We treat the baseline prediction models as a special case of the corresponding cluster-prediction models where we use only one cluster.

5.1 Model Estimation

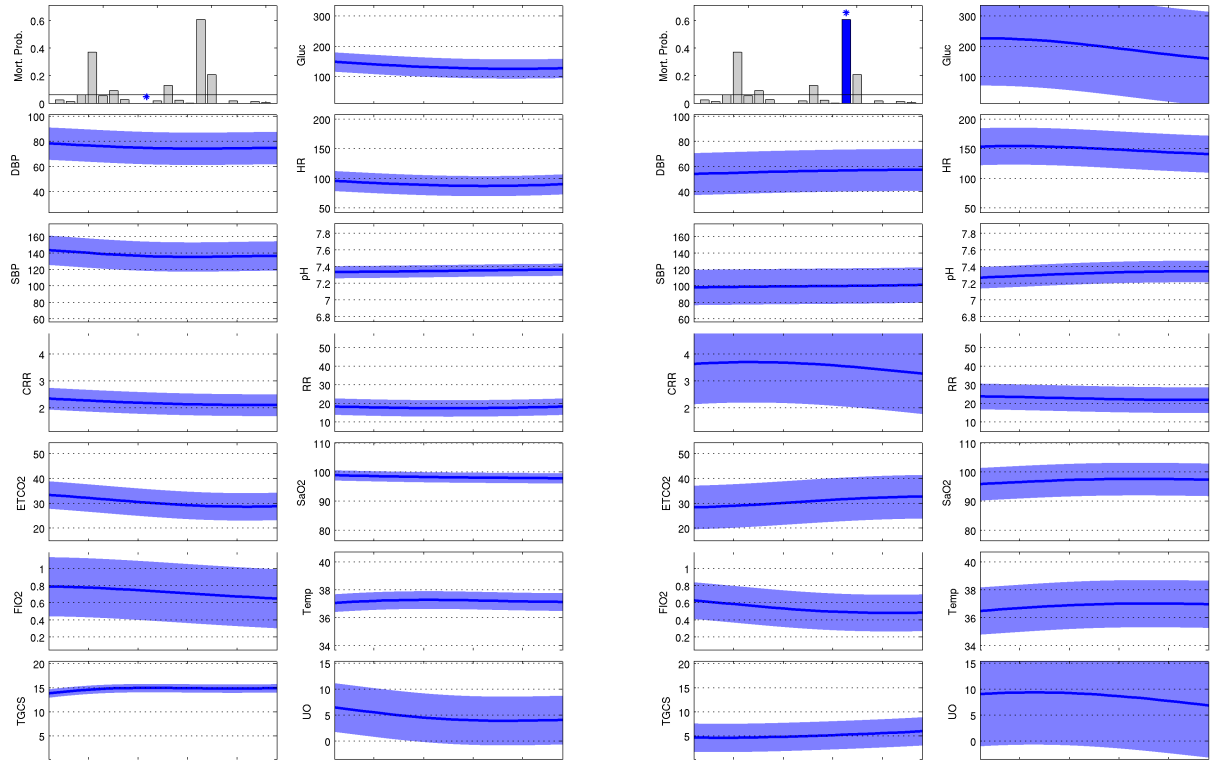
Let y_n be the binary mortality outcome variable, which takes the values 0 (patient lives) and 1 (patient dies). Assume we have a K cluster model. The *Bernoulli Cluster* model has K parameters $\pi_1, ..., \pi_K$ indicating the probability of mortality for a patient episode belonging to each cluster. The model for each cluster is simply $P(Y = 1|Z = k) = \pi_k$. The maximum likelihood estimate given the posterior weighted training sample for cluster k is given below where $q_{nk} = P(Z_n = k|\boldsymbol{x}_n, \boldsymbol{r}_n)$.

$$\pi_k = \frac{\sum_{n=1}^{N} q_{nk} y_n}{\sum_{n=1}^{N} q_{nk}}$$

The *Logistic Cluster* model has a weight vector $\boldsymbol{w}_k$ and a bias b_k for each cluster k. We denote the feature vectors by $\boldsymbol{f}_n$. The logistic regression model for each cluster is given below.

$$P(Y = 1|Z = k, \boldsymbol{f}, \boldsymbol{w}_k, b_k) = \frac{1}{1 + \exp(-(\boldsymbol{w}_k^T \boldsymbol{f} + b_k)}$$

Like the standard unweighted case, the weighted logistic regression estimation problem lacks closed-form solutions for the parameters $\boldsymbol{w}_k$ and b_k. However, the likelihood remains a convex function of the parameters. In the current setting where different clusters may have very different numbers of data cases assigned to them, regularizing the parameter estimates is again very important. We apply a standard $L2$



(a) Lowest Mortality Cluster

(b) Highest Mortality Cluster

Figure 3: Example physiomes corresponding to the components of the 20 cluster model having the lowest and highest probability of mortality.

penalty with strength parameter λ that favors small weight values when there is little data, resulting in less extreme model probabilities. The complete objective function is convex and can be optimized using any numerical optimizer. We apply the limited memory BFGS algorithm [16]. We give the regularized log likelihood function for the weighted case below.

$$\mathcal{L}_k = \sum_{n=1}^{N} q_{nk} \log P(Y_n = y_n | Z_n = k, \boldsymbol{f}_n, \boldsymbol{w}_k, b_k) - \lambda \boldsymbol{w}_k^T \boldsymbol{w}_k$$

5.2 Prediction

Once the models are trained, we make predictions using a similar posterior weighted average. For a novel test case $(\boldsymbol{x}_*, \boldsymbol{r}_*)$, we first compute its posterior probability under each cluster and use these probabilities to combine the per-cluster predictions in a weighted average. The prediction procedure is shown below for both the Bernoulli Cluster and Logistic Cluster models where $q_{*k} = P(Z_* = k | \boldsymbol{x}_*, \boldsymbol{r}_*)$

$$P(Y = 1 | \boldsymbol{x}_*, \boldsymbol{r}_*) = \sum_{k=1}^{K} q_{*k} \pi_k \tag{5.9}$$

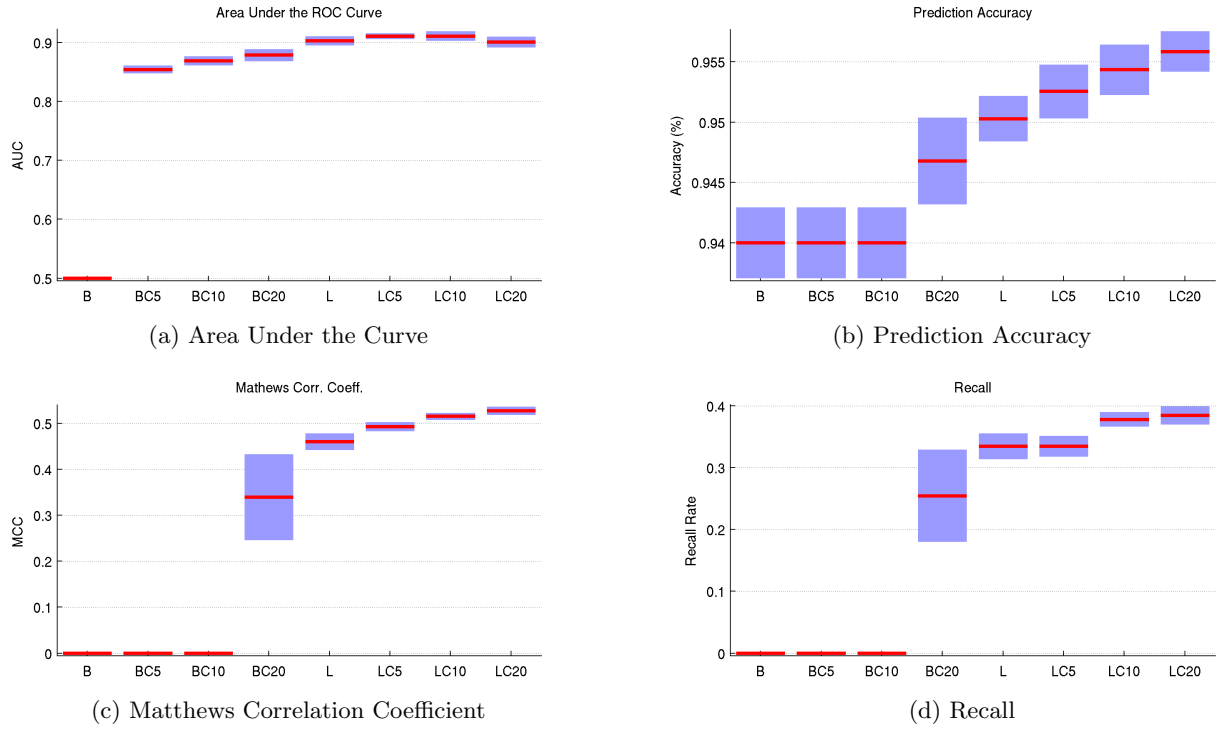
$$P(Y = 1 | \boldsymbol{x}_*, \boldsymbol{r}_*, \boldsymbol{f}_*) = \sum_{k=1}^{K} q_{*k} \frac{1}{1 + \exp(-(\boldsymbol{w}_k^T \boldsymbol{f}_* + b_k)} \tag{5.10}$$

5.3 Empirical Protocols and Performance Measures

We follow a five-fold cross validation procedure identical to that introduced for the clustering experiments. We estimate both the Bernoulli Cluster and Logistic Cluster models given the learned 5, 10 and 20 cluster models estimated previously on each of the five training splits. We train each mortality prediction model on each training split and then evaluate it on the corresponding test split.

We evaluate the predictions made by each model in terms of the area under the ROC curve (AUC), prediction accuracy, Matthews correlation coefficient and recall rate. The AUC score can be interpreted as the probability that a randomly chosen positive data case will be assigned a higher score by the classifier than a randomly chosen negative data case. The predictive accuracy is simply the number of correctly predicted outcomes divided by the total number of episodes (we use 0.5 as the prediction threshold). The Matthews correlation coefficient (MCC) is generally regarded as an accurate performance measure for binary classification problems that is more informative than other measures when the classes are imbalanced [1]. In the current setting, the recall rate is also of great interest since accurate prediction within the small fraction of episodes that result in mortality is highly important.

For all performance measures other than recall, we select the optimal regularization setting λ for the logistic classifier using cross validation on that measure and report the



(a) Area Under the Curve

(b) Prediction Accuracy

(c) Matthews Correlation Coefficient

(d) Recall

Figure 4: Mortality outcome area under the ROC curve, prediction accuracy, Matthews Correlation Coefficient and recall for the Bernoulli prediction models (B, $BC5$, $BC10$, $BC20$) and logistic prediction models (L, $LC5$, $LC10$, $LC20$) as a function of the number of clusters.

corresponding test-set performance value. In the case of recall, we select the regularization setting that maximizes the Matthews correlation coefficient on the validation set and report the corresponding test-set recall rate. Simply maximizing recall is not interesting since we want to achieve good recall while also achieving good performance on a balanced performance measure.

5.4 Mortality Prediction Results

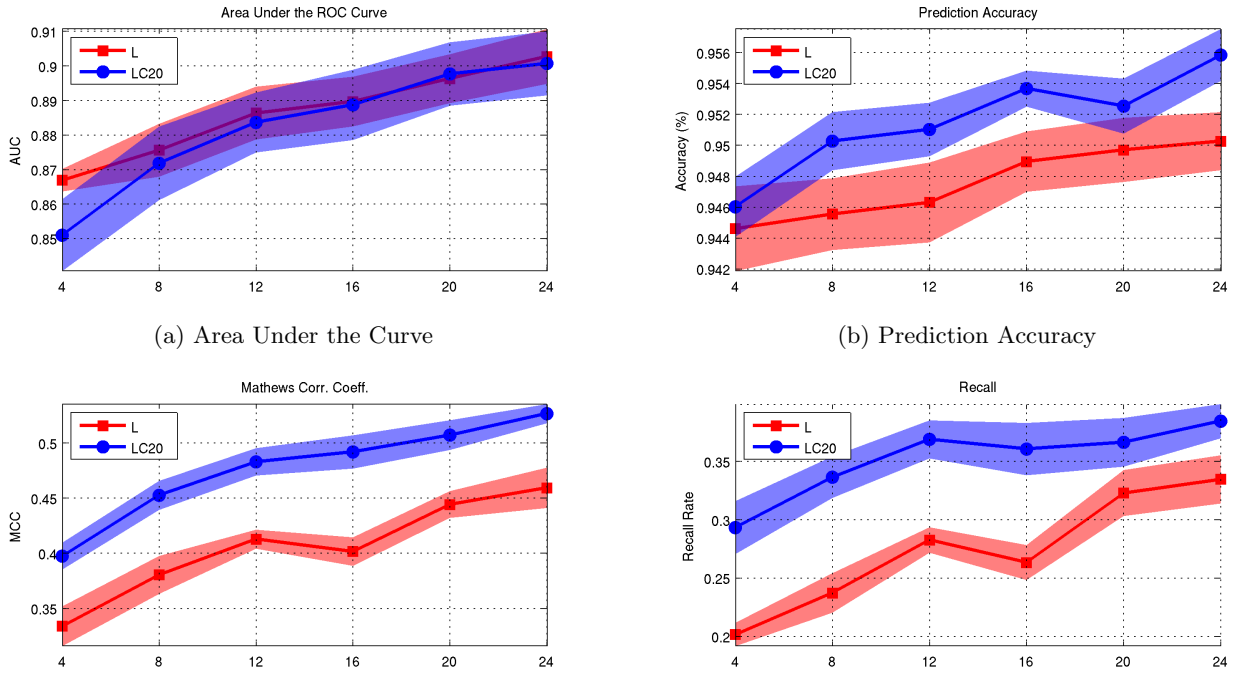
We perform two different mortality prediction experiments. In the first experiment, we compare the predictive performance of the Bernoulli and Logistic prediction models as a function of the number of clusters. Here, the one-cluster models correspond to the baseline Bernoulli and Logistic model, which assume that all of the episodes belong to the same group. We denote the baseline logistic model by L and the baseline Bernoulli model by B. We denote the cluster logistic model with K clusters by LCK and the cluster Bernoulli model with K clusters by BCK.

The results of the first experiment are shown in Figure 4. The lines indicate the mean performance over cross validation splits and the shaded regions correspond to one-standard-error above and below the mean. First, we note that as the number of clusters increases over the range shown here, we obtain significantly improved AUC values using the cluster Bernoulli model. However, the B to $BC10$ models do not predict any positive mortality outcomes when thresholding the mortality probability at the conventional 0.5 level. This yields a flat accuracy curve with respect to the number of clusters, as well as recall and MCC values of exactly 0.

Next, we note that the logistic models L through $LC20$ significantly outperform the corresponding Bernoulli models B through $BC20$ with respect to all performance measures. Importantly, the $LC10$ and $LC20$ models also significantly outperform the baseline logistic model L in terms of accuracy, MCC, and recall, while obtaining an AUC at least as good as that of the baseline logistic model.

In the next experiment, we vary the number of hours of data used when making predictions at test time. The ability to make accurate predictions based on less data is of significant practical interest. We consider making predictions based on the first $T = 4, 8, 12, 16, 20$ and 24 hours of data. We perform inference to compute the probability that a data case belongs to each cluster using only the first T hours of data, treating the remaining time points as missing. We then use the underlying clustering model to impute all the missing observations and extract features as in the previous experiment. We consider only the base logistic model (L) and the 20-cluster logistic model ($LC20$) as it performed the best on the balanced MCC measure in the previous experiment.

We give the results of the second experiment in Figure 5, where the horizontal axes represents the number of hours of data available when making predictions. We see that $LC20$ has significantly better recall than the baseline logistic model over all values of T. The results also show that the MCC value for $LC20$ is significantly better than that of the baseline model over almost all values of T. Importantly, the cluster model achieves these performance gains while yielding MCC and accuracy rates that are better than the baseline logistic model over all T.



(a) Area Under the Curve

(b) Prediction Accuracy

(c) Matthews Correlation Coefficient

(d) Recall

Figure 5: Mortality outcome area under the ROC curve, prediction accuracy, Matthews Correlation Coefficient and recall for the logistic prediction model (L) compared to the 20-cluster logistic prediction model (LC20).

6. DISCUSSION AND FUTURE WORK

In this paper, we have demonstrated a probabilistic clustering model for multidimensional, sparse, uncertain physiological time series data drawn from real-world electronic health care records. This model uses an empirical prior to help overcome issues related to the sparsity of the data. The prior is constructed using a similarity kernel that encourages the mean parameters for each cluster to be smooth with respect to time. We have demonstrated the model's ability to capture patterns of physiology in three ways. First, we visualized the model parameters as time series and showed that the clusters exhibit clear differences in the trajectories of different physiological variables. Second, we showed that the model produces clusters that are associated with large differences in mortality rate. Finally, we demonstrated that the clustering model can be used to construct mortality prediction models that outperform classifiers that treat all episodes as belonging to a single group.

However, the model also has several important limitations. First, the missing at random assumption is certainly violated in cases where a patient dies, resulting in fewer than 24 hours of data. The slight upward trend in TGCS in the high-mortality cluster shown in Figure 3(b) may be indicative of this problem since, under the missing at random assumption, the cluster mean parameters are pulled toward the prior mean (approximately 10 in the case of TGCS) in the absence of data. Second, the discretization of time ignores potentially useful information by grouping observations, possibly decreasing the ability to effectively cluster some patients. Third, the model may also be sensitive to the fact that the start time of each episode is aligned to time of admission and not the onset of the patients underlying condition. Finally, the clustering does not take into account the age-dependence of some variables. These last two issues may result in the unnecessary splitting of some clusters.

There are many possible future directions for this work aimed at overcoming limitations of the current model. We are particularly interested in moving to *mixtures of Gaussian processes* to avoid the discretization of time. Simultaneously, we are very interested in relaxing the assumption that missing values are missing at random. This problem seems to disappear when modeling the raw time series, but it is in fact replaced by a more subtle sample selection bias problem in which a variable's sampling frequency may be related to its underlying value. One avenue toward dealing with these effects is to include an explicit model of the dependency of sampling frequency on value. We are also very interested relaxing the finite time horizon requirement (in which we limit our model to the first T hours), as well as explicitly modeling interventions that have a substantial impact on patient physiology.

In terms of evaluation, we have thus far focused on visualization and mortality prediction. We plan to extend the scope of the evaluation to include different criteria (e.g., goodness of fit tests) and examine the association of the clusters with other phenomena of interest. We are particularly interested in the prediction of clinical diagnoses as well as length-of-stay. We would also like to perform a comparison of the cluster-based predictive model with classic models, such as the *PRISM III* SOI score.

In terms of applications, we are very interested in using cluster-based models to motivate non-trivial notions of *patient similarity* that could be highly useful for exploration of large databases of patient data and for decision support tools driven by retrieval of similar patients. The fact that patient episodes can be different lengths and include different numbers of observations makes the direct computation of standard similarity scores and distances difficult. Discretizing time does not solve this problem since it yields missing data. A natural model-based approach is to learn a large set of clustering models with different parameters and numbers of clusters. We can then measure how frequently each pair of episodes fall into the same cluster to arrive at a similarity score that can be used for patient similarity search, ranking, and information retrieval tasks with massive health data sets.

7. ACKNOWLEDGMENTS

The authors would like to thank Diana MacLean, Chris Mattmann, and Alina Beygelzimer for insightful feedback on this manuscript. This effort was supported in part by the Pacific Institute for the Mathematical Sciences, NLM grant number 1RC1LM010639 and the American Recovery and Reinvestment Act, and the Laura P. and Leland K. Whittier Foundation.

8. REFERENCES

[1] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[2] D. J. Crichton, C. A. Mattmann, A. F. Hart, D. Kale, R. G. Khemani, P. Ross, S. Rubin, P. Veeravatanayothin, A. Braverman, C. Goodale, and R. C. Wetzel. An informatics architecture for the virtual pediatric intensive care unit. In *Proceedings of the Twenty-Fourth IEEE International Symposium on Computer-Based Medical Systems*, pages 1–6, 2011.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B.*, 39(1):1–38, 1977.

[4] Z. Ghahramani and M. I. Jordan. Mixture models for learning from incomplete data. In *Computational learning theory and natural learning systems: Volume IV*, pages 67–85. MIT Press, 1997.

[5] J. J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161, 1979.

[6] D. Kale, A. Hart, C. Mattmann, R. Khemani, P. Ross, P. Vee, J. Terry, R. Wetzel, and D. Crichton. An open source, grid-based software framework for management and sharing of pediatric icu data. In *Proceedings of the Ninth International Conference on Complexity in Acute Illness*, 2010.

[7] E. J. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3):263–286, 2001.

[8] L. Lehman, M. Saeed, G. Moody, and R. Mark. Similarity-based searching in multi-parameter time series databases. In *Computers in Cardiology, 2008*, pages 653 –656, sept. 2008.

[9] S. Leteurtre, A. Duhamel, B. Grandbastien, J. Lacroix, and F. Leclerc. Paediatric logistic organ dysfunction (pelod) score. *The Lancet*, 367(9514):897 – 897, 2006.

[10] S. Leteurtre, A. Martinot, A. Duhamel, F. Gauvin, B. Grandbastien, T. V. Nam, F. Proulx, J. Lacroix, and F. Leclerc. Development of a pediatric multiple organ dysfunction score. *Medical Decision Making*, 19(4):399–410, 1999.

[11] J. Lin and Y. Li. Finding structurally different medical data. In *Proceedings of the Twenty-Second IEEE International Symposium on Computer-Based Medical Systems*, pages 1–8, 2009.

[12] R. Little and D. Rubin. *Statistical analysis with missing data*. Wiley New York, 1987.

[13] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297, 1967.

[14] J. P. Marcin, M. M. Pollack, K. M. Patel, and U. E. Ruttimann. Decision support issues using a physiology based score. *Intensive Care Medicine*, 24:1299–1304, 1998.

[15] G. McLachlan and K. Basford. *Mixture models: Inference and applications to clustering*. Dekker, 1988.

[16] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.

[17] M. M. Pollack, K. M. Patel, and U. E. Ruttimann. Prism III: an updated pediatric risk of mortality score. *Critical Care Medicine*, 24(5):743–752, 1996.

[18] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005.

[19] M. Saeed, G. Lieu, C.and Raber, and R. G. Mark. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Computers in Cardiology*, 29:641–644, 2002.

[20] S. Saria, D. Koller, and A. Penn. Learning individual and population level traits from clinical temporal data. In *the Predictive Models in Personalized Medicine Workshop. Twenty-Fourth Annual Conference on Neural Information Processing Systems*, 2010.

[21] S. Saria, A. K. Rajani, J. Gould, D. Koller, and A. A. Penn. Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants. *Science Translational Medicine*, 2(48):48–65, 2010.

[22] A. Slater, F. Shann, and G. Pearson. Pim2: a revised version of the paediatric index of mortality. *Intensive Care Medicine*, 29:278–285, 2003.

[23] F. Vella. Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources*, 33(1):127–169, 1998.

[24] J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[25] R. C. Wetzel. The virtual pediatric intensive care unit. Practice in the new millennium. *Pediatric Clinics of North America*, 48(3):795–814, 2001.