Using Semi-Parametric Clustering Applied to Electronic Health Record Time Series Data

Suzanne Tamang
Graduate Center
City University of New York
stamang@gc.cuny.edu

Simon Parsons
Graduate Center and Brooklyn College
City University of New York
parsons@brooklyn.cuny.edu

ABSTRACT

We describe a flexible framework for biomedical time series clustering that aims to facilitate the use of temporal information derived from EHRs in a meaningful way. As a case study, we use a dataset indicating the presence of physician ordered glucose tests for a population of hospitalized patients and aim to group individuals with similar disease status. Our approach pairs Hidden Markov Models (HMMs) to abstract variable length temporal information, with non-parametric spectral clustering to reveal inherent group structure. We focus on systematically comparing the performance of our approach with two alternative clustering methods that use various time series statistics instead of HMM based temporal features. Intrinsic evaluation of cluster quality shows a dramatic improvement using the HMM based feature set, generating clusters that indicate more than 90% of patients are similar to members of their own cluster, and distinct from patients in neighboring clusters.

1. INTRODUCTION

Temporal mining is a special case of data mining that seeks to address the methodological issues presented by real-world databases that are temporal in nature. Tasks typical of these methods include: data characterization and comparison, clustering analysis, classification, association rules, pattern analysis and trend analysis [8].

Although Electronic Health Records (EHRs) provide opportunities to uncover important patterns and discover new knowledge in medicine and healthcare, significant challenges exist when processing the huge volumes of temporal data they contain. We aim to develop a robust time series clustering method to facilitate temporal mining in EHR repositories. As a case study, we cluster patients into characteristic groups using time series data for patients with one or more physician ordered glucose test in their EHR. Our preliminary work moves towards establishing a high-performance, temporal mining framework that can be used to reveal structure among a population of patients with shared clinical characteristics, and to develop an effective temporal mining method that can easily generalize to new temporal analysis problems. Figure 1 shows an overview of the approach.

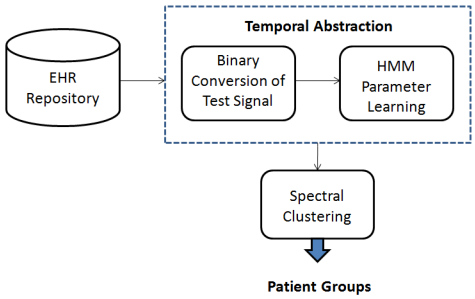

Figure 1: Temporal Mining Framework

The rest of this paper is structured as follows. Section 2 discusses previous efforts made by researchers for time series analysis and the main theoretical or algorithmic ideas in our paper. Section 3 describes our research methodology. Section 4 presents the experimental results and Section 5 then concludes the paper and sketches our future work.

2. BACKGROUND

2.1 Temporal Abstraction

Temporal abstraction techniques are used to provide a description of a time series when it is infeasible to process data in the raw form. In the context of clustering, this is mostly due to the size of the dataset, but other problems result from processing sequences that are non-uniformly sampled, variable in length, highly heterogenous or incomplete.

2.1.1 Hidden Markov Models (HMMs)

The parametric assumption of Markov models is useful for summarizing temporal information. Most previous work identifies a set of k-HMM components that can be used to describe a time series dataset [7][9] but other extensions of Markov based abstraction methods are noted in the research literature [1].

An HMM articulates the patient's sequence using a set of model parameters, $\lambda = \{A, B, \pi\}$, over a set of N hidden states with M possible emissions. Since a patient's disease status is not directly observable, our model assumes that a patient is in one of three disease states (N=3): *stable*, *moderate*, or *unstable*. Also, that the state emissions, M, which

KDD-DMH '11, August 21, 2011, San Diego, California, USA.

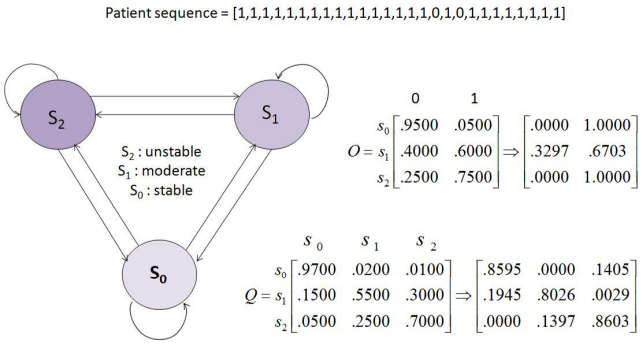

Figure 2: HMM example

in our model consist of the daily *presence* "1" or *absence* "0" of a medical test (M=2), function as an indicter for an unseen disease related phenomena.

For time series abstraction, we train an HMM model for each patient glucose time series. Using a three state model, Figure 2 shows an example of how parameters are estimated from a patient's sequence represented as an emission, or 0/1 observation vector. The HMM parameter A is an NxN matrix that represents the probability of moving from the current state to the next state (e.g., transitioning to an unstable state tomorrow, when stable today) and B is an NxM matrix indicating the probability that a glucose test was ordered in each of the N states. Since the model parameters A and B are not known, the Baum Welch algorithm is used to compute the parameter estimates. In our example, the matrices O and Q show the initial transition and emission probabilities that were used, π, and the posterior estimates, A and B, that are calculated for the example sequence using Baum Welsh, and later flattened to provide a feature vector for clustering. A more detailed description of HMMs is provided by Rabiner [5].

2.2 Clustering

Cluster analysis is attractive in that methods can be used to find patterns in the data that are not predicted by the researcher's current knowledge or pre-conceptions. The goal of clustering algorithms is to divide data into clusters that are meaningful or useful, and improving existing techniques has been the focus of considerable research in machine learning. Often, clustering is an exploratory process that is used in the preliminary investigation of relatively unexplored data set.

2.2.1 Time Series Clustering

Related time series clustering work using parametric Markov models to represent time series data has been shown to produce high performance clustering results in other domains. For example, recent work by work by Hu et al. demonstrates the effectiveness of pairing HMMs with hierarchical clustering [2]. Other work [3] shows that performing clustering with a non-parametric spectral methods allows for the clustering algorithm to be as agnostic as possible in regard to the shape of clusters. Although our work deviates in implementation details, we also use HMMs to abstract variable length time series in a concise representation, and spectral clustering methods to impose as few constraints on cluster formation as possible.

Table 1: Descriptive Statistics for Study Population

Feature	Average	SD	Min	Max
lor	1012	8	1000	1025
numT	24	37	1	497
lenGap	540.11	284.41	15.00	1023.00
fDays	0.0234	0.0367	0.0001	0.4849
hMeas	0.0958	0.0906	0.0078	0.6927
hTime	4.6117	1.7762	-0.5000	7.8186

3. METHODS

To determine the performance of HMM temporal abstraction paired with non-parametric spectral methods, hereafter referred to as *semi-parametric clustering*, we compare the results of our approach with two alternative clustering algorithms that use summary statistics of the glucose time series instead of HMM based abstraction to represent temporal phenomena. Time series features that were available to the alternative clustering approaches included: entropy of the measurement sequence, entropy of the time between tests, length of record, duration of the longest time gap between tests, fraction of days tested and the total tests.

Although one of the alternative clustering approaches used spectral methods to produce clustering assignments, the semi-parametric clustering approach is unique in that it uses HMM parameter estimates as spectral clustering features. Many varieties of spectral clustering algorithms exist and in this work we use a method first proposed by Ng et al. [4], which normalizes the Laplacian affinity matrix before eigenvalue decomposition and selection of k largest eigenvalues.

3.1 Data

Our study uses de-identified time series data that was obtained from a population of patients hospitalized at New York-Presbyterian Hospital (NYPH) with at least one physician ordered glucose test indicated in their EHR. The glucose time series presents methodological challenges in that it is non-uniformly sampled in time, variable in length, and incomplete. To demonstrate the feasibility of our clustering approach, we selected all patients with a time series length in the range of 1000 to 1025 days for our study. The dataset that was generated resulted in 1024 patient 0/1 measurement sequences. Table 1 shows patient level summary statistics for the total study population including: length of record (lor), total number of tests (numT), fraction of days for which glucose tests were ordered (fDays), entropy of the measurement sequence (hMeas) and entropy of the time differences between tests (hTime).

3.2 Model Selection Criteria

To select the best model for the different approaches, we first looked for a dramatic drop in the clustering procedure's objective function. If no dramatic drop could be observed, or more than one distinct angle resulted, Bayesian information criteria was used to select among competing models.

For semi-parametric clustering our criteria was used to determine the best value of k. Since some of the temporal features available to the alternative clustering approaches and noted in Table 1 are highly correlated, we applied the criteria using different variable combinations and values of k.

3.3 Cluster Evaluation

Intrinsic validation of the patient clusters was performed using the silhouette validation technique, which is based on the comparison of a cluster's *compactness* and its *separation* from other clusters [6]. For each patient, $a(i)$ is the average dissimilarity of patient i to all patients in its respective cluster. We then find the average dissimilarity of patient i with patients of another cluster, repeating for all clusters that patient i is not a member of. The cluster with the lowest average dissimilarity is the "neighboring cluster" of patient i and indicated by $b(i)$. The silhouette value is defined as:

$$s(i) = \frac{(b(i) - a(i))}{\max\ (a(i), b(i))}$$

and the average $s(i)$ of a clustering assignment is a measure of how tightly patients are grouped into their respective clusters and how distinct clusters are with respect to each other. When $s(i)$ is equal or above .6, patient i can be considered appropriately clustered. A value close to -1 indicates that a patient would have been more appropriately assigned to the neighboring cluster, $b(i)$, and a value close to 1 indicates that individuals in the patient's respective cluster are very similar and that the cluster is distinct from other clusters.

4. RESULTS

Using the model selection criteria described in Section 3.2, we determined the model settings for the alternative approaches, k-Means and spectral clustering without HMM abstraction. Both used three clustering features: entropy of the measurement sequence, length of longest gap, and the number of visits; however, the two methods differed in the value of k.

The semi-parametric clustering method differs from the alternative approaches in the use of HMMs for time series abstraction. To select between two candidate models, k=4 and k=9, information criteria was used, indicating that that the simpler model was was slightly better. Since the relative difference were minor, we chose to retain both models for future evaluation by a domain expert.

4.1 Intrinsic Evaluation

To compare the quality of clusters generated by the different clustering approaches, we used the silhouette validation technique described in Section 3.3. Silhouette values are a heuristic commonly used to assess the goodness of clusters, providing information on both inter-cluster compactness and the level of distinction between different clusters. Experimental result for k-Means, spectral clustering without HMM abstraction (SC), and our semi-parametric clustering approach that pairs HMMs and spectral clustering (HMM+SC) is shown in Table 2. Each cell corresponds to the percent of patients that meet various silhouette thresholds (SV); specifically, the percent of patients with values equal to or greater than 0.6 (the minimum value associated with a good clustering assignment), 0.7, 0.8 and 0.9. Also, this table shows the percent of patients with negative silhouette values, which is a strong indication that they were incorrectly assigned.

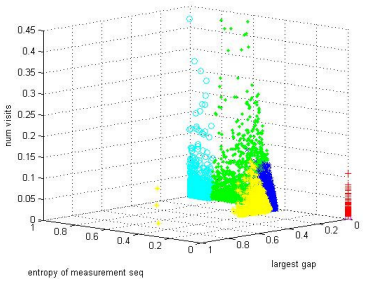

Figure 3: Entropy Based Clustering ($k = 5$)

k-Means and Spectral Clustering without HMM Abstraction: Our results show that k-means clustering resulted in worst the performance with under 20% of patients organized into useful groups. Additionally, 12% of patients showed negative silhouette values, suggesting that they are more appropriately assigned the neighboring cluster. The spectral clustering method without HMM abstraction performed comparatively better, reporting 63% of patients were grouped in an appropriate cluster; however, overall, both clustering algorithms that used summary statistics about the time series as clustering features failed to group patients into useful clustering assignments.

Figure 3 plots the results from spectral clustering without HMM abstraction and provides some intuition as to why clustering results using time series summary statistics are generally poor. That is, model selection methods identify the entropy of the 0/1 measurement sequence as an important feature for clustering, but the shape of the entropy function appears to distort the feature space. Also, k-means assumes that clusters are Gaussian and Figure 3 shows that the results for spectral clustering without HMM abstraction provides strong evidence for the contrary. Spectral methods relax the Gaussian assumption; however, the feature space is still problematic and the algorithm cannot effectively assign patients into clusters with distinct boundaries.

Semi-parametric Clustering: When compared to the alternative clustering methods, the semi-parametric approach that pairs HMM abstraction to represent each patient's time series with spectral clustering (HMM+SC) shows notable improvements. Two candidate models resulted from the model selection procedure and we report silhouette values for both. For the candidate model k=4, over 90% of patients had a silhouette value equal or greater than 0.60, indicating a useful clustering assignment. The alternative model, k=9, shows that 84% of patients has silhouette values equal to or greater that 0.6. As noted previously, silhouette values close to 1 indicate a strong level of cluster compactness and distinction from other clusters. For the k=4 model 75% of patients report a value of 0.9 or greater, suggesting that semi-parametric clustering can be used to produce

Table 2: Comparison of Clustering Results

SV	k-Means	SC	HMM+SC,4	HMM+SC,9
.60	18.73%	62.73%	93.56%	84.10%
.70	10.73%	48.88%	91.32%	79.22%
.80	3.12%	27.80%	86.05%	67.61%
.90	0.98%	13.46%	74.83%	40.88%
neg	12.34%	1.91%	0.20%	-

high-quality patient clusters. Figures 4 show the silhouettes for HMM+SC where k=4.

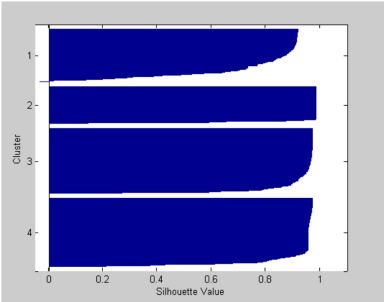

Figure 4: Silhouette Values for $k = 4$

5. CONCLUSIONS AND FUTURE WORK

To analyze large volumes of temporal EHR data, methods that can scale, demonstrate high-performance, and provide the flexibility needed to generalize to new problems will prove the most useful to researchers. However, many problems are posed by the nature of EHR data, which does not fit the canonical time series framework and is often non-uniformly sampled in time, highly heterogenous and incomplete, making temporal analysis of patient data challenging.

Using a semi-parametric temporal clustering framework, we aim to develop a robust method for clustering patient time series data that can generalize to other temporal clustering problems. Our approach pairs the parametric assumptions of HMMs to abstract patient time series with a non-parametric spectral clustering technique to reveal inherent group structure. To assess the effectiveness of our approach, we compare experimental results with two alternative clustering methods, k-Means and spectral clustering without HMM based temporal clustering features. Based on intrinsic evaluation, our results show that clusters formed by the alternative approaches were not sufficiently distinct, and that a non-trivial number of patients were assigned to the wrong cluster by k-means. Although entropy of the time series is an informative feature to describe a patient's glucose time series, the shape of the entropy function distorts the clustering space, and results in the formation of clusters without salient boundaries. In contrast to the alternative approaches, silhouette values generated for the semi-parametric clustering approach strongly indicate that our method can be used to generate high quality clusters and capture the temporal and spectral aspects of patient EHR data.

Although this work demonstrates the effectiveness of HMMs as a abstraction method for patient time series, and improved performance using a non-parametric clustering approach, testing on a larger-scale and validation by domain experts is essential to determine the clinical significance of our results. Directions for future work include extrinsic evaluation of our clusters to more accurately evaluate performance and improving clustering results in the semi-supervised setting. Recent clustering work in other domains where a full set of labeled examples are difficult to generate, and there is not enough data to build a classifier, has shown that using some labeled data to place constraints on the clustering model can significantly improve clustering results.

6. REFERENCES

[1] R. Datta, J. Hu, and B. Ray. Sequence mining for business analytics: Building project taxonomies for resource demand forecasting. In *Proceeding of the 2008 conference on Applications of Data Mining in E-Business and Finance*, pages 133–141, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[2] J. Hu, B. K. Ray, and L. Han. An interweaved hmm/dtw approach to robust time series clustering. In *ICPR (3)'06*, pages 145–148, 2006.

[3] T. Jebara, Y. Song, and K. Thadani. Spectral clustering and embedding with hidden markov models. In *Proceedings of the 18th European conference on Machine Learning*, ECML '07, pages 164–175, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

[5] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[6] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65, 1987.

[7] S. P. Shah, K.-J. Cheung, N. A. Johnson, G. Alain, R. D. Gascoyne, D. E. Horsman, R. T. Ng, and K. P. Murphy. Model-based clustering of array cgh data. *Bioinformatics*, 25:i30–i38, June 2009.

[8] T. Warren Liao. Clustering of time series data-a survey. *Pattern Recogn.*, 38:1857–1874, November 2005.

[9] Y. Zeng and J. Garcia-Frias. A novel hmm-based clustering algorithm for the analysis of gene expression time-course data. *Comput. Stat. Data Anal.*, 50:2472–2494, May 2006.