Similarity measuring between patient traces for clinical pathway analysis

Similarity Measuring between Patient Traces for Clinical Pathway Analysis

Zhengxing Huang, Xudong Lu, and Huilong Duan*

College of Biomedical Engineering and Instrument Science of Zhejiang University, The Key Laboratory of Biomedical Engineering, Ministry of Education, China
duanhl@zju.edu.cn

Abstract. Clinical pathways leave traces, described as activity sequences with regard to a mixture of various latent treatment behaviors. Measuring similarities between patient traces can profitably be exploited further as a basis for providing insights into the pathways, and complementing existing techniques of clinical pathway analysis, which mainly focus on looking at aggregated data seen from an external perspective. In this paper, a probabilistic graphical model, i.e., Latent Dirichlet Allocation, is employed to discover latent treatment behaviors of patient traces for clinical pathways such that similarities of pairwise patient traces can be measured based on their underlying behavioral topical features. The presented method, as a basis for further tasks in clinical pathway analysis, are evaluated via a real-world data-set collected from a Chinese hospital.

1 Introduction

Clinical pathway analysis (CPA) has experienced increased attention over the years due to its importance to health-care management in general and to its usefulness for capturing the actionable knowledge and interesting insights to administrate, automate, and schedule the best practice for individual patients in clinical pathways [1]. A carefully inspection of patient traces can support health-care organizations to analyze and improve clinical pathways. By measuring similarities between patient traces, it can be useful to health-care organizations for a number of reasons including better overall clinical pathway management and maintenance [2].

In order to measure similarities between patient traces, it is a common technique to provide a measure of distance in the features' space, e.g., to compute similarity primarily by using activity sequences of patient traces. Traditional techniques of sequence similarity measures are focused on direct matching between sequences applying commonly the classical distance concepts. They may not be appropriate to measure similarities between patient traces for clinical pathways.

In this study, we employ a probabilistic graphical model, i.e., Latent Dirichlet Allocation (LDA) [3], to measure similarities between patient traces for clinical

* Corresponding author.

N. Peek, R. Marín Morales, and M. Peleg (Eds.): AIME 2013, LNAI 7885, pp. 268–272, 2013.

pathways. The assumption made is that the possibly treatment behaviors of patient traces in clinical pathways may be represented by a relatively small number of simple and common behavioral topics, which can be combined with the original patient traces to measure similarities between traces. We use real-life data from Zhejiang Huzhou Central Hospital of China to evaluate the proposed method.

2 Method

In this study, we assume that it is possible to sequentially record various kinds of clinical activities in clinical pathways. In general, hospital information systems record such information. To introduce the patient trace representation model and our similarity measure method, we first define the following concept.

Definition 1. *Let $\mathcal{A}$ be the set of clinical activities. A patient trace is a non-empty sequence of clinical activities performed on a particular patient, i.e., $c = \langle a_1, a_2, \ldots, a_n \rangle$, where $a_i \in \mathcal{A}$ $(1 \leq i \leq n)$ is a particular clinical activity. For convenience, let $c(i)$ be the ith clinical activity in the trace. A patient trace repository R is a multi-set of patient traces.*

In general, LDA helps to explain the behavioral similarity of patient traces by grouping clinical activities into unobserved sets. A mixture of these sets then constitutes the observable patient trace. The generative process of LDA is as follows. For each patient trace c, a mixture of topic proportion $\theta_c \sim Dir(\alpha)$ is sampled from a Dirichlet distribution parameterized by the hyperparameter α. Each clinical activity a in a trace is generated by first sampling a topic t from a multinomial distribution $t \sim Mult(\theta)$, and then sampling $a \sim Mult(\phi_t)$ also from a multinomial distribution. Given a treatment behavioral topic t, each $\phi_t \sim Dir(\beta)$ is sampled from a Dirichlet distribution parameterized by β. In LDA, each patient trace c is a mixture of topics represented by θ_c and each topic t is a distribution over all activities represented by $\phi_{t,a} = Pr(a|t)$.

Using this generative model, the treatment behavioral topic assignments for clinical activities can be calculated based on the current topic assignment of all the other clinical activity positions. More specifically, the topic assignment is sampled from:

$$Pr(t_i = t|t_{\neg i}, c) = \frac{n^a_{t,\neg i} + \beta}{\sum_{b \in A} n^b_t + \beta|A|} \frac{n^t_{c,\neg i} + \alpha}{\sum_{j \in K} n^{t_j}_c + \alpha K} \tag{1}$$

where $t_i = t$ represents the assignment of the ith occurrence to topic t, $t_{\neg i}$ represents all treatment behavioral topics assignments not including the ith occurrence, K is the number of topics, $|A|$ is the number of clinical activities, $n^a_{t,\neg i}$ is the number of times activity a is assigned to topic t, not including the current instance, and $n^t_{c,\neg i}$ is the number of times topic t is assigned to the patient trace c, not including the current instance.

From these count matrices, we can estimate the topic-activity distribution θ and trace-topic distribution ϕ by,

$$\theta_{t,a} = \frac{n_t^a + \beta}{\sum_{b \in A} n_t^b + \beta|A|}, \ \phi_{c,t} = \frac{n_c^t + \alpha}{\sum_{t \in T} n_c^t + \alpha K} \tag{2}$$

Exact inference in LDA is generally intractable. In particular, we use Gibbs sampling to estimate the parameters of the LDA model. Once we have learned the model parameters, we can measure the similarity between patient traces. In particular, for a specific trace c in the repository R, we obtain the topic distribution $\overrightarrow{\theta_c} = \{\hat{\theta}_{c,t_1}, \hat{\theta}_{c,t_2}, \cdots, \hat{\theta}_{c,t_K}\}$, where each $\hat{\theta}_{c,t_i}$ is the posterior estimate of θ_{c,t_i} for the treatment behavioral topic t_i $(1 \leq i \leq K)$. Upon this, we are able to calculate the similarity between two traces c and c^* $(c, c^* \in R)$ as follows:

$$sim(c, c^*) = \frac{\sum_{t \in T} \hat{\theta}_{c,t} \times \hat{\theta}_{c^*,t}}{\sqrt{\sum_{t \in T} \hat{\theta}_{c,t}^2} \sqrt{\sum_{t \in T} \hat{\theta}_{c^*,t}^2}} \tag{3}$$

To illustrate the feasibility of the proposed approach, we present a specific application, i.e., patient trace clustering, based on similarities between patient traces. Patient trace clustering helps reveal the underlying characteristics and commonalities among a large collection of traces. The information extracted by clustering can also facilitate subsequent analysis, for instance, to extract common treatment patterns of execution in the traces, or speed up trace indexing and anomaly detection. A reasonable similarity measure $sim(c, c^{'})$ is critical for the patient trace clustering. The objective of the clustering methods that work on similarity measure function is to maximize the intra cluster similarities and minimize the inter cluster similarity. In this study, we adopted a hierarchical micro-clustering algorithm to generate partitions of patient traces in the repository.

3 Case Study

The experimental data set was extracted from Zhejiang Huzhou Central hospital of China. In the experiments, we build a specific patient trace repository of clinical pathways of several specific types of cancer, i.e., branchial lung cancer, colon cancer, rectal cancer, breast cancer, and gastric cancer, from the system. The collected data is from 2007/08 to 2009/09. In detail, there are 258 traces, 11028 clinical activities with 266 activity types. In the experiments, we conducted topic analysis for the experimental repository using LDA with the different number of treatment behavioral topics $(K = 1, 2, \cdots, 20)$. The Dirichlet prior α and β of LDA are set to 0.2 and 0.1. The number of iterations of Gibbs sampling is set to 10000. In addition, to expand the number of trials when we construct the LDA model, we adopt a fivefold cross-validation strategy.

As shown below, the presented method is evaluated by a specific application, i.e., patient trace clustering. In particular, we compare the presented LDA-based similarity measure with the traditional edit-distance-based similarity measure

[4]. In the following experiments, we refer to LDA-based similarity measure with K-topic model ($K = 1, 2, \cdots, 20$) as LDA-K, and edit-distance-based similarity measure as ED.

The benchmark clusters are identified from the experimental repository. In particular, we use the first diagnosis code to category patient traces. As mentioned above, 5 categories, i.e., bronchial lung cancer, colon cancer, rectal cancer, breast cancer, and gastric cancer, are extracted from the repository, which can be used as benchmark clusters for evaluating the overall performance of clustering. For evaluation on patient trace clustering, we measure "$F_{0.5}$" to calculate the accuracy of the system on a per-trace basis and then build a global score for all patient traces in the repository.

Using the benchmark clusters, we can evaluate clustering performance on $F_{0.5}$. In particular, by taking the maximum value of $F_{0.5}$ (among different merging thresholds ε from 0.0 to 0.4), we compare the performance of ED and LDA-K ($K = 1, 2, 3, \cdots, 20$). As shown in Figure 1, when the number of topics is larger than a particular value ($K \geq 8$), the $F_{0.5}$ is quite stable. Certainly, $k \approx 8$ is probably the suitable number of topics for the experimental patient trace repository.

Now we study the impact of the parameter ε on both the experimental results, where ε is the merging threshold in the clustering step. We vary the value of ε from 0.0 to 0.4. Figure 2 shows the results of ED and LDA-8 (using the 8-topic model). From Figure 2, we can see that LDA-8 can provide significant improvement over ED. The maximum value of $F_{0.5}$ of LDA-8 is 0.6422, which is nearly 56% better than ED (0.1044). Note that when margining threshold is zero, each patient trace is classified into a specific cluster. That explains why both curves have the same starting value of $F_{0.5}$ shown in Figure 2. In addition, the inclusion of latent topics increases similarity among patient traces. As a result, when merging threshold is small, LDA-8 does not show an advantage over ED. When merging threshold increases, LDA-8 obtains better results on $F_{0.5}$ than ED, while the latter remains stable regardless of the value of ε. In particular, LDA-8 provides the most significant improvements when ε is 0.15. Note that we can always obtain better results with LDA-8 except $\varepsilon = 0$ in comparison with

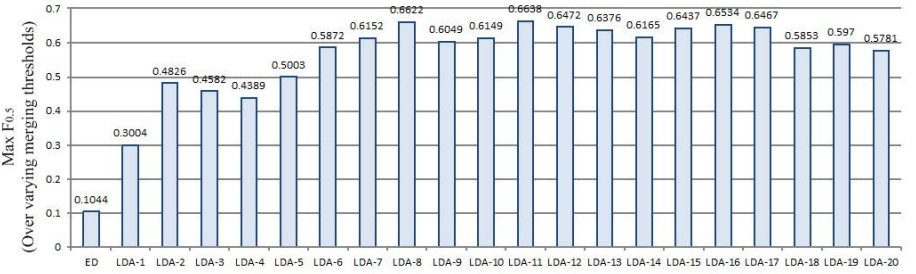

Fig. 1. Performance of clustering using ED and LDA with different latent treatment behavioral topic models on the experiment repository

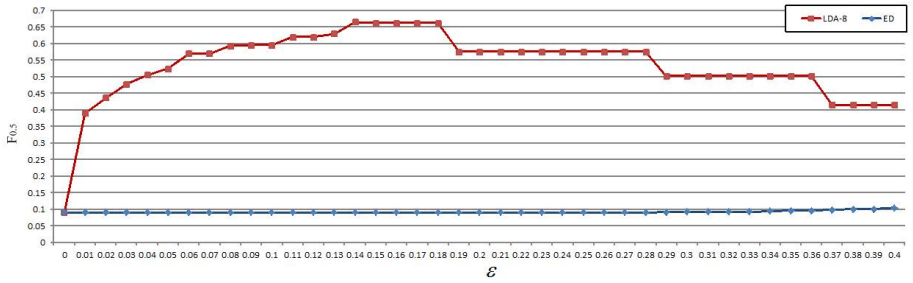

Fig. 2. The comparison between ED and LDA-8 on patient trace clustering

ED. It indicates that the treatment behavioral features have more influences on the similarity measure and subsequent analysis (e.g., patient trace clustering) than the sequential order of clinical activities of the traces.

4 Conclusion

In this paper, we have introduced a new method of measuring the similarities between patient traces for clinical pathways, which can profitably be exploited as a basis for further tasks of CPA, e.g., critical/essential treatment behaviors can be detected, analyzed, and optimized based on the topic analysis presented in this study, association rules between recognized anomalies and patient states can be derived, etc. We will address these tasks by exploiting the potential of the proposed similarity measure between patient traces and its applications, as a crucial advantage over traditional techniques for clinical pathway analysis and optimization.

Acknowledgment. This work was supported by the National Nature Science Foundation of China under Grant No 81101126.

References

1. Huang, Z., Lu, X., Duan, H.: Latent treatment topic discovery for clinical pathways. Journal of Medical Systems 37(2), 1–10 (2013)
2. Combi, C., Gozzi, M., Oliboni, B., Juarez, J.M., Marin, R.: Temporal similarity measures for querying clinical workflows. Artificial Intelligence in Medicine 46(1), 37–54 (2009)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Joural of Machine Learning Research 3, 993–1022 (2003)
4. Gusfield, D.: Algorithms on strings, trees and sequences, Computer Science and Computational Biology. Cambridge University (1997)