Analysis of Medical Treatments Using Data Mining Techniques

Xin Xiao, Silvia Chiusano
Dipartimento di Automatica e Informatica, Politecnico di Torino - Torino, Italy Email:
{xin.xiao,silvia.chiusano}@polito.it

***Abstract*—Since in health care systems the amount of data is continuously growing, data mining techniques can be applied to analyse these large collections and gain interesting insights. However, some critical issues should be properly addressed. For example, data collections on patient treatments are usually characterized by an inherent sparseness and variable distribution, due to the large variety of possible treatments performed by patients affected by a given disease. To effectively extract interesting knowledge from such collections, we present a framework coupling a *clustering* and a *classification* algorithm. The clustering approach is named *multiple-level* because we apply the clustering algorithm in a multiple-level fashion, by focusing on different dataset portions and locally identifying groups of patients with similar profile and examination history. The classification algorithm is used to characterize the discovered clusters. This paper also describes future research issues and possible developments of the proposed framework.**

***Index Terms*—Data mining, cluster analysis, classification analysis, medical records, patient examination history.**

I. Introduction

Nowadays, large amount of medical data, storing the patient medical history, is collected during health care. The analysis of these medical data collections is a challenging task for health care systems since a huge amount of interesting knowledge can be automatically mined to effectively support both physicians and health care organizations.

Data mining techniques [1], which focus on studying effective and efficient algorithms to transform large amounts of data into useful knowledge, have been widely exploited on medical data by analyzing different pathologies or different aspects of the same disease. For example, previous studies addressed food analysis [2] and investigated risk factors associated with diabetes and pre-diabetes [3], while current issues in medicine knowledge discovery are discussed in [4].

Analysing real world health care data collections may impose new challenges. These collections can have *large volume* and *high dimensionality* due to the large cardinality of patient records and the variety of medical treatments usually adopted for a given pathology. In addition, they are usually characterized by a *variable data distribution* and *inherent sparseness*. Consequently, innovative data mining approaches are needed to efficiently gain interesting insights from such collections.

In this paper we present a framework to discover, in a patient data collection with a variable distribution, cohesive and well-separated groups of patients with a similar profile (i.e., patient age and gender) and examination history (given by the set of examinations performed by patients). The framework couples a clustering approach (named “multiple-level clustering”) for cluster set computation, and a classification algorithm used to both characterize the cluster content and measure the effectiveness of the clustering process. Health care organizations can exploit the discovered knowledge for example to check the coherence between the adopted treatments and existing medical guidelines for a given disease, as well as enrich the existing guidelines or assess new ones.

The paper is organized as follows. The framework is presented in Section II, while future research issues and possible developments of the framework are discussed in Section III.

II. The proposed data analysis framework

The presented framework is depicted in Figure 1 and summarized in the following.

To deal with the inherent sparseness and variable distribution of patient data collections, a density-based *multiple-level clustering* approach is adopted in the framework. We named the approach “multiple-level” because it performs multiple runs over the considered data collection. This strategy aims at progressively partitioning the initial data collection into (quite) homogeneous subsets, thus easing the computation of cohesive clusters on each of them. Specifically, at each iteration a different dataset portion is analyzed, and clusters are locally identified on it.

A novel distance measure has been defined to cluster patients according to the three aspects characterizing them, i.e., patient age, gender and examination history. For the data representation, the patient examination history tailored to the Vector Space Model (VSM) is used, and the examination frequency is weighted using the TF-IDF (Term Frequency (TF) - Inverse Document Frequency (IDF)) score [1]. TF-IDF has been used in text mining to analyse document collections, with the aim of weighting the relevance of words in documents. In our context, we used this approach to weight the relevance of examinations in the patient examination history, highlighting peculiar examinations for each patient. The examination frequency can vary significantly from standard tests to specific examinations used to diagnose disease complications. TF-IDF allows focusing on examinations specific for each patient and discarding examinations done by most patients.

The discovered cluster set is evaluated through the Silhouette index [1], and with the support of domain experts to describe the cluster content from a medical perspective. Specifically, a class label is assigned to each cluster.

Starting from the labeled cluster set, a *classification model* is created both to characterize the content of clusters and measure the effectiveness of the clustering process. This model can be used to automatically assign a new patient to a given class based on her/his profile and examination history. In addition, when the adopted classification algorithm provides a readable model (e.g., decision trees [1]), this model can give useful insights to domain experts on some peculiar properties characterizing patients in each class (in terms of gender, age and undergone examinations).

As a first attempt, the proposed framework has been used in [5] to analyse a real dataset of diabetic patients provided by an Italian Local Health Center. The DBSCAN algorithm [1] has been adopted for the multiple-level clustering approach, while decision trees to compute the classification model.

Results showed that the multiple-level clustering approach progressively discovers clusters containing patients with increasing disease severity. For example, clusters computed in the first iteration of the approach contain patients mainly undergoing routine tests to monitor diabetes conditions. Instead, clusters computed in the next iterations contain patients tested using an increasing number of examinations to diagnose several diabetes complications. The cluster set is characterized by good Silhouette values, and the classification model computed from it is very accurate (above 90% of accuracy), with good precision and recall values for most class labels.

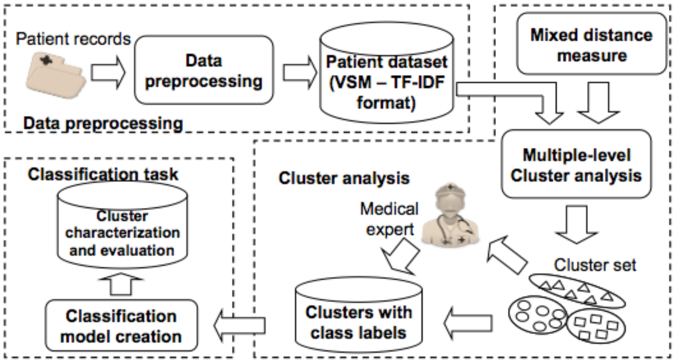
Patient records
Data preprocessing
Patient dataset (VSM – TF-IDF format)
Data preprocessing
Mixed distance measure
Multiple-level Cluster analysis
Classification task
Cluster analysis
Medical expert
Cluster characterization and evaluation
Cluster set
Classification model creation
Clusters with class labels

Fig. 1. The proposed framework

III. Future work

This section describes some research directions for the analysis of medical data, and specifically possible developments of the framework presented in Section II.

(i) Evaluating alternative clustering and classification algorithms. Different clustering and classification algorithms can be selected for integration in the proposed framework. Based on the target application scenario, the proper algorithms can be adopted by considering different issues as final number of clusters and average cluster size, classification model accuracy and readability, and computational cost.

(ii) Analysing additional information on patient treatments. Besides patient examination history, additional aspects of the medical treatments such as prescribed drugs can also be considered. This analysis can help discovering correlations between prescribed drugs and disease complications, as well as detecting and preventing drug misusing. The information on prescribed drugs can be used to characterize patient clusters computed using the framework presented in Section II, or to drive the clustering process for discovering groups of patients with similar examination histories and drug therapies.

(iii) Using data taxonomy in the data analysis process. Taxonomies can be used to generalize examinations and drugs into their corresponding categories. They can be used to drive the process of clustering patient data, or to reduce the data dimensionality problem by considering medical data described at different abstraction levels.

(iv) Considering the temporal dimension in the patient examination history. Since medical data includes frequently occurring temporal patterns in patient records, the temporal dimension in medical data analysis can be a critical issue. Temporal mining approaches can be used to discover clusters of patients who have similar temporal relations among the examinations. The temporal analysis can help to identify examination pathways commonly followed by patients, and potentially check and improve predefined guidelines.

References

[1] Pang-Ning T. and Steinbach M. and Kumar V., *Introduction to Data Mining*. Addison-Wesley, 2006.

[2] M. Phanich, P. Pholkul, and S. Phimoltares, “Food recommendation system using clustering analysis for diabetic patients,” in *IEEE International Conference on Information Science and Applications (ICISA)*, 2010, pp. 1–8.

[3] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, “Comparison of three data mining models for predicting diabetes or prediabetes by risk factors,” *The Kaohsiung Journal of Medical Sciences*, no. 0, 2012.

[4] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend,” *Expert Systems with Applications*, vol. 41, pp. 4434–4463, 2014.

[5] G. Bruno, T. Cerquitelli, S. Chiusano, and X. Xiao, “A clustering-based approach to analyse examinations for diabetic patients,” in *IEEE International Conference on Healthcare Informatics*, 2014, pp. 45–50.