Applied Soft Computing 14 (2014) 31–37

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A medical procedure-based patient grouping method for an emergency department

M. Xu[a,1], T.C. Wong[a,*], K.S. Chin[a,b,2]

[a] Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China
[b] Centre for Systems Informatics Engineering, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 15 March 2013
Received in revised form
25 September 2013
Accepted 28 September 2013
Available online 6 October 2013

Keywords:
Healthcare
Patient grouping
Emergency department
Medical procedures
Self-organizing map

ABSTRACT

Variability and unpredictability are typical characteristics of complex systems such as emergency department (ED) where the patient demand is high and patient conditions are diverse. To tackle the uncertain nature of ED and improve the resource management, it is beneficial to group patients with common features. This paper aims to use self-organizing map (SOM), *k*-means, and hierarchical methods to group patients based on their medical procedures and make comparisons among these methods. It can be reasonably assumed that the medical procedures received by the patients are directly associated with ED resource consumption. Different grouping techniques are compared using a validity index and the resulting groups are distinctive in the length of treatment (LOT) of patients and their presenting complaints. This paper also discusses how the resulting patient groups can be used to enhance the ED resource planning, as well as to redesign the ED charging policy.

1. Introduction

Emergency department (ED) plays a critical role in most of the healthcare systems because it is the frontline of acute care and the main route of admission to hospital. The complexity of an ED system is featured by diverse patient conditions, fluctuating patient demands, and manpower variation. In Hong Kong, the EDs have been prone to violating the ED service pledges such as the maximum waiting time and length of stay due to large patient demand and persistent ED manpower shortage. Grouping patients with similar characteristics would be one of the feasible alternatives to reduce the complexity and uncertainty in the ED management, and hence ensure the delivery of quality and stable services to the ED patients. Some previous researchers have been dedicated to patient grouping but most of them are not successful in clustering patients in consideration of resource consumption [1]. Three commonly used patient grouping criteria are introduced below.

Casemix is a very common approach of patient grouping based on diagnosis-related grouping (DRG). As its name suggests, it only considers the patient clinical diagnosis categories [1]. The basic premise of Casemix is that patients with similar DRGs will consume a similar amount of resources. In many countries, local governments are using Casemix to fund public hospitals. However, people who against this approach may argue that the DRG method is built on the assumption of "average consumption" which could limit its predictive capabilities of identifying total resource consumption. For example, the elderly and frail patients would suffer under the DRG-based payment because they usually consume more ED resources, but only average resource level is assigned to them [2]. Moreover, the DRG method was originally developed to manage hospital patients. Its effectiveness under a more dynamic and pressing ED system has not been tested by any study.

Length of stay (LOS) grouping suggests that LOS is a good proxy measure of resource consumption when the direct measurement is difficult and costly [3,4]. Many studies have adopted Gaussian mixture models to group hospital patients based on their LOS [4–6]. However, the appropriateness of using LOS as proxy to resource consumption is questionable when being applied to ED. In ED, most of the patients usually spend a large proportion of their LOS in the waiting for the first doctor consultation, during which little ED resources are consumed. Therefore, the time between the first consultation and disposition (discharge, admission, dead), or hereinafter known as length of treatment (LOT), would be a better measure of ED resource consumption than LOS.

Patient pathway grouping is based on patient physical movements within a healthcare facility. Isken and Rajagopalan [7] have proposed an approach of grouping patients according to the

* Corresponding author at: Department of Systems Engineering and Engineering Management, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong, China. Tel.: +852 3442 6155; fax: +852 3442 0172.
E-mail addresses: maixu@student.cityu.edu.hk (M. Xu), andywtc@graduate.hku.hk (T.C. Wong), mekschin@cityu.edu.hk (K.S. Chin).
[1] Tel.: +852 3442 2564; fax: +852 3442 0172.
[2] Tel.: +852 3442 8306; fax: +852 3442 0172.

1568-4946/$ – see front matter
http://dx.doi.org/10.1016/j.asoc.2013.09.022

pathway they took within hospital area. Maruster et al. [8] have suggested an alternative approach of grouping patients with respect to the logistic perspective of treatment. Takakuwa and Shiozaki [9] have identified over 70 patterns of patient flows for 9 patient categories in a simulation project for an ED in Japan. One problem of this grouping method is that patients with different medical needs may share the same pathway. Therefore, the ED staff assigned to each pathway must possess a wide range of skills to meet various medical needs. Also, it is quite challenging to optimize the arrangement of material resources for different pathways.

Three grouping methods above cannot be explicitly linked up with the "actual" resource consumption in ED. Given the uncertain nature of the modern ED system, grouping patients requiring common medical procedures is worthy of investigation if one wants to modularize the ED management for better decision making especially under scarce resources. It is, thus, the objective of this paper to use a data-driven method to group patient with similar pattern of resource consumption. As the name indicates, the data-driven method makes solutions totally based on the data, and it does not involve any human knowledge which might bias the grouping result.

The organization of the paper is as follows: Section 2 introduces several grouping methods and describes the study workflow. Section 3 describes the implementation of the proposed method using the real data followed by result discussion. Section 4 discusses how the patient grouping can be helpful to the ED manpower planning and charging policy. Section 5 concludes this study together with future work.

2. The proposed methodology

Jain et al. [10] have provided a comprehensive survey of existing grouping techniques and some important applications. Nearly all the grouping algorithms aim to construct clusters with minimal intra-group diversity and maximal inter-group distinction. In this study, the patient grouping is a non-supervised learning process because no patient groups can be pre-defined from any specific ED knowledge. Therefore, three unsupervised grouping techniques are considered here, namely, hierarchical methods, k-means, and self-organizing maps (SOM).

The hierarchical methods can be further classified into two types, namely agglomerative and divisive. The former starts with n clusters, where n is the number of observations. The distance between observations is calculated, and the two closest points are merged into a cluster. The number of groups is set by the analyst in advance before the merging process begins. On the contrary, the divisive hierarchical methods start from one cluster, that is the entire population, and the closest points are merged into new clusters. The drawback of hierarchical methods is that it cannot handle large amount of data and it is easily affected by the outliers. Moreover, the non-recovery characteristic of these methods means once an observation is assigned to a cluster, it cannot be moved at all [11].k-means is a popular non-hierarchical grouping technique. It is characterized by simple algorithm and fast convergence especially for high-dimensional problems. In essence, it is a method of partitioning a given set of n data points into k groups in D-dimensional Euclidean space R^D. The partitioning in the space is usually based on Euclidean distance. k-means can have higher grouping accuracy if the number of groups is appropriately determined in advance. The process of k-means algorithm is summarized in four steps (Table 1).

Table 1
Steps of k-means clustering algorithm.

Step	Description
1	Selection k number of initial centroids $\{z_0, z_1, \ldots, z_{k-1}\}$ randomly from data points $\{x_0, x_1, \ldots, x_{n-1}\}$.
2	Assign a data point x_i, $0 \le i \le n$, to the group C_j, $0 \le j \le k$, if and only if $\|\|x_i - z_j\|\| < \|\|x_i - z_p\|\|$ for $j \neq p$, where $0 \le p < k$.
3	Compute new centroid as $z_j^* = \frac{1}{\left\|C_j\right\|}\sum_{x_i \in C_j} x_i$, where z_j^* is the new value of z_j.
4	If less or no change of the centroids, then stop, otherwise continue to the step 3.

SOM [12] is a type of artificial neural network based on unsupervised learning algorithm, and it is a popular grouping method in solving real problems [15–17,21]. It can provide a two-dimensional visual presentation of high-dimensional data. It works well with large data size and multiple input variables. The non-linear nature of SOM overcomes some of the limitations of the hierarchical methods given its superiority of robustness against missing data. Similar to k-means, SOM does not assume prior distribution of data. The pseudo-code of SOM learning is shown in Fig. 1.

The comparisons between three grouping techniques have been made in some previous studies. Mangiameli et al. [16] have presented an extensive work demonstrating that SOM outperformed seven hierarchical methods in terms of accuracy and robustness using 252 data sets with various levels of imperfections including dispersion, outliers, and non-uniform cluster densities. Same conclusion was drawn by Waller and Kaiser [17] claiming SOM outperformed hierarchical methods using artificial data sets. Kuo et al. [11] have made a comprehensive comparison between these three methods. They have found that both SOM and k-means are advantageous in handling large dataset, but the former is able to generate repeatable results due to its non-sensitivity to the starting points (number of groups and centroids). SOM has been extensively applied in different fields such as Engineering, medicine, tourism, marketing, economics, physics, and chemistry as summarized by Oja et al. [18].

The existing literature however is not conclusive over the best grouping technique. In fact, the superiority of a method largely depends on the data [19,20]. Thus, this paper intends to compare the grouping result from hierarchical methods (single-link, complete-link, and average-link), SOM, and k-means using the real data of a local ED. The best grouping method is selected by the validity index. Because the grouping process is non-supervised, it is impossible to calculate the grouping accuracy against the pre-defined results from specific ED knowledge. Instead, the grouping results will be interpreted in the medical sense.

Thus, the key steps of the proposed medical procedure-based grouping method are briefly summarized below.

Data collection: We retrieve all the patient records during the study period. Besides the medical procedures, each patient record also includes information of demographics and triage category.

Data preparation: We clean and preprocess the data into desirable format for analysis.

Grouping methods: We apply hierarchy methods, SOM, and k-means to the patient data.

Method selection: We select the most desirable grouping method based on the validity index.

Validation: First, we determine if the grouping results are clinically reasonable and able to differentiate among patient groups. Second, we investigate if each group is distinct in terms of resource consumption. Here, LOT is used as a proxy of resource consumption in ED as previously discussed.

3. Implementation and results

3.1. Data collection, cleansing, and preprocessing

The data contains 2452 records of Category 3 and 4 patients in a random week of 2012. It was collected manually from the archived