Discriminating Patient Length of Stay by Using the k-Means Clustering Algorithm. An Application to a Surgical Database

Florin GORUNESCU

Department of Mathematics, Biostatistics and Computer Science,
University of Medicine and Pharmacy of Craiova,
Craiova, Romania
fgorun@rdslink.ro

Abstract. The aim of this paper is to use the k-means clustering algorithm for a surgical database in order to obtain an optimum length of stay segmentation by automatic means. Thus, it can be provided the hospital staff with a real picture of the number of patients against different LOS in that hospital, enabling good decisions making and thus optimizing the health care costs.
Keywords: length of stay, k-means clustering
Math. Subject Classification 2000: 91C20

1 Introduction

The health care issues worldwide known encompass the following two aspects:

- **Health care systems are complex.** According to the World Health Report 2000 - Health systems: improving performance (WHO, 2000), modern health care systems involve: good health, responsiveness to the expectations of the population and fair financial contribution. Thus, the health care industry, now the nation's second-largest employer in the USA, has had an increasingly important impact on the economy. Public concern about the management and delivery of health care services is near the top of the national agenda, and the factors and policies influencing health care systems in around the world are a natural subject for many researchers in the domain of the modeling and optimization.
- **Increasing demand and escalating costs.** Nowadays, health care spending continues to rise at the fastest rate in history. Experts agree that all health care systems are riddled with inefficiencies, excessive administrative expenses, inflated prices, poor management, and inappropriate care, waste and fraud. These problems significantly increase the cost of medical care and health insurance for employers and workers and affect the security of families.

One practical solution to partially solve these problems is represented by modeling and optimizing the patient flow through hospitals, thus reducing costs and increasing the medical service quality.

The *patient length of stay* (LOS) represents the duration of time a patient spends in hospital. A classic metric for gauging the success of different health care policies is (hospital) length of stay. LOS has become an important measurement used to control costs, commonly used as an indication of the quality of care rendered, and is a common outcome variable used to compare the performance between hospitals. The efficient utilization of resources has become a priority in the health care environment today (e.g. reducing LOS purportedly yields large cost savings). LOS may offer a reliable method of evaluating efficiency of bed utilization and may complement mortality ratios in assessing effectiveness of care.

The aim of this paper is to automatically group patients according to their LOS, by using the k-means clustering algorithm. This segmentation of patients according to their LOS will provide both better allocation and scheduling of resources and improvement of the management of patients. Previous work [1] used a Gaussian mixture modelling (GMM) to predict the optimal number of groups in order to develop a patient classification and prediction methodology to identify groups of patients according to LOS.

2 The dataset

The dataset used in this study concerns a number of 7,723 (surgical) patients, the records consist of: gender, age, date of admission and discharge, public or private patient, emergency/planned, diagnosis (MDC). The distribution of data (LOS -days) is displayed below (Fig. 1).

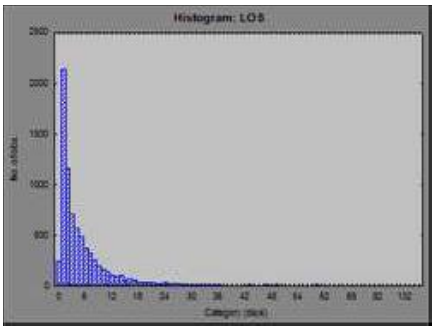

Fig. 1. Surgical LOS distribution

The main statistical details concerning the surgical dataset are displayed in Table 1, while Fig. 2 visualizes this result.

Table 1. Statistical details (surgical LOS)

	Valid N	Mean	Confidence -95%	Confidence +95%	Min.	Max.	Std. Dev.
LOS	7723	5.85	5.64	6.05	0.00	228.00	9.25

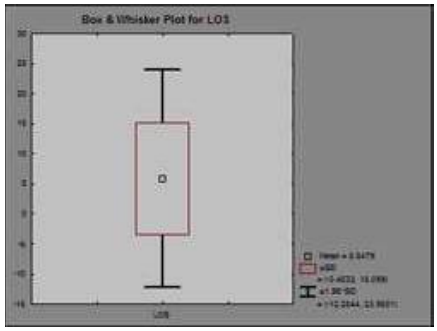

Fig. 2. Box & plots for LOS

The data normality has been checked up using both the standard Kolmogorov-Smirnov & Lilliefors and Shapiro-Wilk tests and the conclusion drawn shows an expected non-Gaussian behavior (the null hypothesis that the distribution is normal is rejected with the p-level = 0.01/0.00) -see Fig. 3 below.

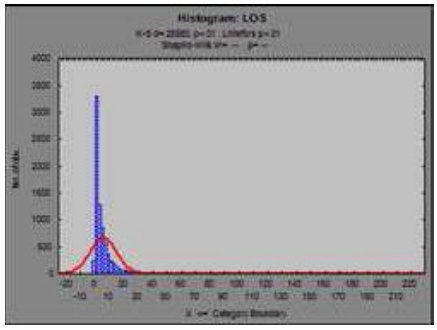

Fig. 3. Testing normality (surgical LOS)

3 k-means clustering

The *k-means algorithm* is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. Basically, k-means is an algorithm to cluster n objects based on attributes into k partitions, $k < n$. The main idea is to define k *centroids*, one for each cluster, and to populate the corresponding clusters with the nearest items to them. A main issue is how to choose the initial centroids, because different locations cause different results. Thus, a good choice is to place them as much as possible far away from each other. Another idea is to select them in a way that they are already initially close to large quantities of points. In this paper the former option has been chosen. The next step is to take each point belonging to a given dataset and associate it to the nearest centroid. When no point is pending, the first step is completed and the recalculation of k new centroids of the clusters resulting from the previous step is needed. After the new k centroids have been computed, a new reallocation has to be done between the same dataset points and the nearest new centroids. As a result of this iterative procedure, the k centroids will change their location step by step until no more changes are done. Finally, the algorithm aims to minimize the objective function, given by the squared error function:

$E_{rr} = \sum_{j=1}^{k} \sum_{i=1}^{n} \parallel x_i^{(j)} - c_j \parallel^2$

and the centroid c_j is defined by:

$c_j = \frac{1}{n_j} \sum_{i \in S_j} x_i^{(j)}$

where $\parallel \cdot \parallel$ is a chosen distance measure (usually the Euclidian distance) between a data point $x_i^{(j)}$ and the centroid c_j of the j-th cluster S_j.

The basic k-means algorithm consists of the following steps [2], [3]:

1. Place k points into the space represented by the objects that are being clustered. These points represent the initial cluster centroids.
2. Assign each object to the cluster that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the (new) k centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move.

Remark. The k-means algorithm does not necessarily find the most optimal configuration corresponding to the global objective function minimum, the result strongly depending on the value of k. Unfortunately, there is no general theoretical science to find the optimal number of clusters for any given data set. A simple approach to evaluate the appropriateness of the classification consists in comparing the within-cluster variability (small if the classification is good) to the between-cluster variability (large if the classification is good).

4 LOS segmentation results

The aim of this study was to identify the (near) optimum number of groups concerning the LOS for the surgical dataset. We have used the k-means clustering algorithm (initial centroids chosen to maximize distances between them) and performed a "What-If" analysis by varying the number of clusters from 3 to 6. The statistical results of this analysis are shown in Table 2.

Table 2. Descriptive statistics for different number of clusters (3 to 6 clusters))

No. of clusters	Clusters	Mean (SD) days	No. of patients (percentage)	95% confidence interval for average LOS (days)
3	C1	3.36 (2.78)	6753 (87.5%)	(3.29, 3.43)
	C2	18.97 (6.63)	876 (11.3%)	(18.53, 19.41)
	C3	62.06 (29.46)	94 (1.2%)	(56.1, 68.0)
4	C1	3.21 (2.58)	6619 (85.7%)	(3.15, 3.27)
	C2	17.48 (6.12)	988 (12.8%)	(17.1, 17.86)
	C3	53.12 (15.37)	112 (1.4%)	(50.27, 55.97)
	C4	178.25 (39.30)	4 (0.01%)	(139.74, 216.76)
5	C1	2.63 (1.92)	6002 (77.7%)	(2.58, 2.68)
	C2	11.54 (3.15)	1303 (16.9%)	(11.37, 11.71)
	C3	26.90 (5.86)	337 (4.4%)	(26.27, 27.53)
	C4	59.39 (14.67)	77 (0.9%)	(56.11, 62.67)
	C5	178.25 (39.30)	4 (0.1%)	(139.74, 216.76)
6	C1	2.63 (1.92)	6002 (77.7%)	(2.58, 2.68)
	C2	11.54 (3.15)	1303 (16.9%)	(11.37, 11.71)
	C3	26.90 (5.86)	337 (4.4%)	(26.27, 27.53)
	C4	59.39 (14.67)	77 (0.9%)	(56.11, 62.67)
	C5	161.67 (25.81)	3 (0.009%)	(132.46, 190.88)
	C6	228 (0.00)	1 (0.001%)	(228, 228)

From Table 1 we see that:

- Larger the (mean) LOS for a certain cluster, scattered (sparser) the 95% confidence interval. Recall that the 95% confidence interval for the sample mean represents the range of values which contains the true population mean with probability 0.95.
- Smaller the (mean) LOS for a certain cluster, smaller the corresponding standard deviation (SD). Recall that a large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.
- As an overall conclusion: shorter the LOS, better prediction for the average LOS.

The goal of the k-means clustering algorithm is to classify objects into a user-specified number of clusters. The main drawback of this clustering procedure consists in the fact that there is no rule to choose the optimal number of clusters. To evaluate the appropriateness of the data segmentation and thus the number of clusters, we performed an analysis of variances, comparing the within-cluster variability -small if the classification is good- to the between-cluster variability -large if the classification is good. Consequently, using this metaheuristic, it is obtained that a (near) optimal number of cluster range between 4 and 5 (p-level < 0.01) -see the graph below (Fig. 4), showing the between-cluster variability (the top curve) against the within-cluster variability (the bottom curve).

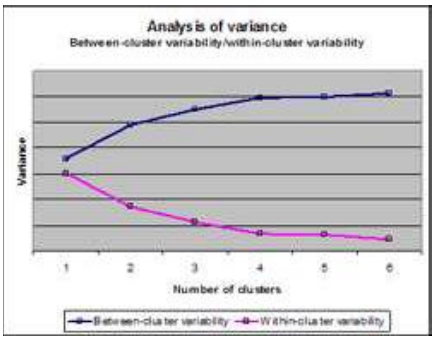

Fig. 4. Analysis of variances (intra/inter cluster)

As is displayed above, for 4 and 5 clusters there is equilibrium between the two trends, denoting a balance between the inter-cluster and intra-cluster variability. From analyzing the two options -4 or 5 clusters- the following conclusions are to be drawn:

- In the case of four clusters we have:
  - About 86% of patients stay for an average (rounded) LOS of 3 days in hospital;
  - About 13% of patients stay for an average LOS of 17-18 days in hospital;
  - About 1% of patients stay for an average LOS > 53 days in hospital;
- In the case of five clusters we have:
  - About 78% of patients stay for an average LOS of 2-3 days in hospital;
  - About 17% of patients stay for an average LOS of 11-12 days in hospital;
  - About 4.4% of patients stay for an average (rounded) LOS of 27 days in hospital;
  - About 0.5% of patients stay for an average (rounded) LOS > 60 days in hospital.

References

1. **Moody, J., Darken, C.J.**, Fast learning in networks of locally-tuned processing units. Neural Computation 1(2), 281–294 (1989)
2. **El-Darzi, E., Abbi, R., Vasilakis, Ch., Gorunescu, F., Gorunescu, M.**, Length of stay-based clustering methods for patient grouping. In: 2th International Health and Social Care Modelling Conference-HSCM2008, 18 - 20 March 2008, Portrush, Northern Ireland, (2008)
3. **MacQueen, J.B.**, Some Methods for classification and Analysis of Multivariate Observations. In: Proc. 5-th Berkeley Symposium on Mathematical Statistics and Probability, (1) pp. 281–297. Berkeley, University of California Press, (1967)