Annals of University of Craiova, Math. Comp. Sci. Ser.
Volume 36(2), 2009, Pages 79–84
ISSN: 1223-6934

Patients length of stay grouping using the hierarchical clustering algorithm

Smaranda Belciug

Abstract. Patient management is a very challenging problem of the health care system. Grouping patients according to their Length of Stay (LoS) in the hospital leads to a better planning of bed allocation, and patient admission and discharge. The aim of this paper is to statistically evaluate and cluster the data using the agglomerative hierarchical clustering algorithm.

Hospitals nowadays confront with a very serious matter regarding resource management. A bad management can lead to various problems such as: an under-provision of hospital beds means building-up waiting lists and places a lot of stress on both the hospital system and patients (when insufficient medical beds are provided to meet demand, emergency medical patients spill over into surgical beds; consequently, surgical waiting lists increase as planned admissions are postponed), or an over-provision of hospital beds is wasteful of scarce resources.
The solution to this problem is patient management. Patient management means generating, planning, organizing, and administering medical and nursing care and services for patients. As a direct consequence of individuality, patients typically differ in a number of medical, physical and socio-economic characteristics, (age, severity of illness, complications, speed of recovery, pain thresholds). Groups of patients with health care needs, whether they are rushed into hospital as emergencies or suffer from a particular disease can be considered as groups of patients with similar needs. These groups are heterogeneous, thus requiring more detailed modeling for classification. Health care needs and corresponding resources vary from patient to patient, thus a great stress is placed on the health care system in planning and managing efficiently the resources.
Similar to bed occupancy groupings and phase-type approaches, classification algorithms can also be used to discern the different types of patients based on their LoS in hospital. Unsupervised classification algorithms aim to derive or distinguish between different groups and can therefore be used to enhance understanding and make predictions in the presence of large volumes of historical data.
In this paper the agglomerative hierarchical clustering has been used to group the patients into clusters and also a statistical evaluation of the database has been provided.

1. Patient management in stroke medicine

A stroke is the rapidly developing loss of brain function(s) due to a disturbance in the blood supply to the brain. As a result, the affected area of the brain is unable to function, leading to inability to move one or more limbs on one side of the body, inability to understand or formulate speech or inability to see one side of the visual field.

Ideally, people who have had a stroke are admitted to a "stroke unit", a ward or dedicated area in hospital staffed by nurses and therapists with experience in stroke treatment. It has been shown that people admitted to a unit have a higher chance of surviving than those admitted elsewhere in hospital, even if they are being cared for by doctors with experience in stroke.
Stroke rehabilitation is the process by which patients with disabling strokes undergo treatment to help them return to normal life as much as possible by regaining and relearning the skills of everyday living[2], [3], [4]. It also aims to help the survivor understand and adapt to difficulties, prevent secondary complications and educate family members to play a supporting role. A rehabilitation team is usually multidisciplinary as it involves staff with different skills working together to help the patient. These include nursing staff, physiotherapy, occupational therapy, speech and language therapy, and usually a physician trained in rehabilitation medicine. As it can be seen managing patients that suffered from one or more strokes is a very important and difficult task.

2. Hierarchical clustering

Data clustering methods can be hierarchical. These algorithms can be either agglomerative - "bottom - up" - , or divisive - "top-down". The algorithms are clustering data using previously established clusters, thus building successive clusters.
Agglomerative algorithms begin with each data being a separate cluster and afterwards merge them into larger clusters. Divisive algorithms begin with a whole set and proceed to divide it into smaller clusters.
Hierarchical clustering creates a tree structure, also known as a dendrogram, which represents the hierarchy of clusters. The root of this dendrogram consists of a single cluster that contains all the observations, and the leaves correspond to individual observations.
The steps of the hierarchical agglomerative clustering algorithm are:
1. Compute the proximity matrix
2. Let each data point be a cluster
3. Repeat
4. Merge the two closest clusters
5. Update the proximity matrix
6. Until only a single cluster remains

The strengths of hierarchical clustering involve:
- Do not have to assume any particular number of clusters, since any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level;
- The clusters may correspond to meaningful taxonomies.

3. Dataset

We tested and applied the above method on a stroke dataset, which albeit not recently collected, contain data that are still typically stored by hospital computerized systems. The Stroke dataset originates from the English Hospital Episode Statistics (HES) database and concerns all finished consultant episodes of stroke patients, aged 65 and over discharged from all English hospitals between April 1st 1994 and March 31st 1995 (105,765 episodes) [5]. The variables describing each spell include age, admission method, admission source, main specialty, gender, regional health authority

of treatment, district health authority of treatment, number of diagnosis codes, season, weekend admission, diagnosis coded using international classification of disease (ICD) codes, discharge method and discharge destination. A patient spell is qualified as stroke if it contains a stroke related diagnosis code anywhere in the diagnostic chain (stroke related diagnoses are between codes 430 and 438 in the International Classification of Diseases, Injuries and Causes of Death-Revision 9, ICD-9). No information that identified individual patients was supplied. The average LOS is 14 days (standard deviation 52 days), the median is seven days and the range is between zero and 4,907 days.

4. Results

The main statistic parameters included in this analysis are the following:
- Mean;
- Median;
- Confidence interval;
- Standard deviation (SD).

Statistical parameters					
Stroke	Mean	Median	Conf.Int. 95 % (mean)	SD	Min / Max
LoS (days)	13.10	7.00	(12.73, 13.46)	48.16	0 / 4906

Table 1 shows that there is a large enough variance of the data (SD = 48.16), that is the LoS of the stroke patients is very spread (see also the gap between minimum and maximum LoS). This is a good reason to analyze the LoS dataset to find patterns, mainly provided by the cluster analysis. Moreover, there is a significant difference between the mean and the median of data, proving the existence of outliers (very large number of days, even for a small number of patients).
As visualization methods, the histograms and box & whiskers techniques were used. Histograms are used in the following way:
- They show the distribution of values of a single variable;
- They divide the values into bins and show a bar plot of the number of objects in each bin;
- The height of each bar indicates the number of objects;
- The shape of histogram depends on the number of bins.
The box and whisker plot will summarize each variable by three components:
- A central square to indicate central tendency or location (e.g. mean);
- A box to indicate variability around this central tendency;
- Whiskers around the box to indicate the range of the variable.
Note. If data follows the normal distribution, then mean $\pm$ 1.96·SD between whiskers represents the 95% confidence interval and, therefore, the corresponding data should fall within the whiskers.

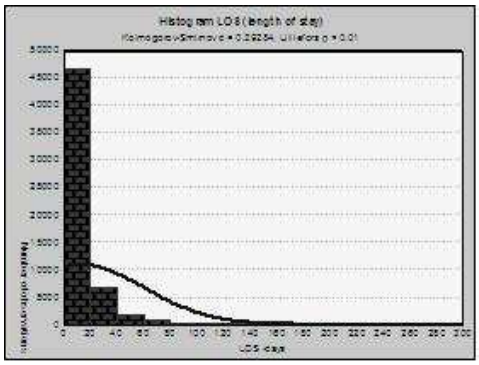

Figure 1. LOS histogram

Observe that the normal distribution graph (half Gauss bell) is depicted together with the corresponding histogram, visually showing that the stroke dataset is far from a Gaussian distribution. From Fig. 1 we see that the majority of observations (56,148) belongs to the interval (0, 100], followed by zero days in hospital (8,882), and (100, 200] -378 observations.

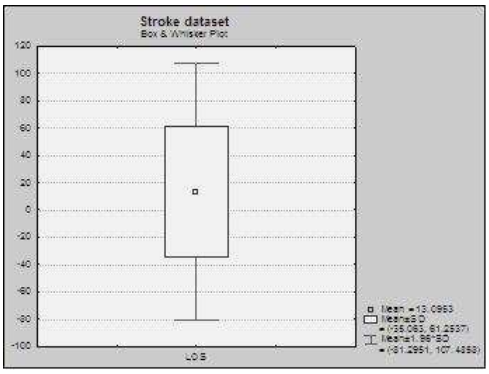

Figure 2. Box & Whiskers plot -LOS

From Fig. 2 we see that the majority of data (mean ± SD) lies in the interval [0, 61], while the data belonging to mean ± 1.96·SD lies in the interval [0, 107]. This observation strengthens the idea that the large majority of stroke patients spent short time in hospital.

Many types of statistical analysis are based on the assumption that the data are normally distributed. Although simple descriptive statistics such as skewness and kurtosis can provide some relevant statistical information, more precise information can be obtained by performing tests of normality to determine the probability that the sample came from a normally distributed population of observations. Thus, we have used both the *Kolmogorov-Smirnov & Lilliefors* test, which is applicable when the mean and the standard deviation are computed from the actual data, and the *Shapiro-Wilk W* test (Altman 1991).

The *Kolmogorov-Smirnov D test* for normality is based on the maximum difference between the sample cumulative distribution and the hypothesized cumulative distribution. If the corresponding D statistic is significant, then the hypothesis that the respective distribution is normal should be rejected. Ideally, the probability values that are reported are valid when the mean and standard deviation of the normal distribution are known *a priori* and not estimated from the data. However, those

parameters are usually computed from the actual data and thus the *Lilliefors* significance level should be considered instead.[1]

Table 2. Normality test (Kolmogorov-Smirnov)

Normality test (Kolmogorov-Smirnov)			
Variable	Kolmogorov-Smirnov	Skewness	Kurtosis
	K - S max D / Lilliefors p	47.95	3354.54
LoS - stroke	0.39 / < 0.01	-	-

Table 2 shows that the stroke dataset it is not normally distributed. Technically, *skewness* measures the deviation of the distribution from symmetry. If the skewness is clearly different from 0, then that distribution is asymmetrical, while normal distributions are perfectly symmetrical. *Kurtosis* measures the "peakedness" of a distribution. If the *kurtosis* is clearly different than 0, then the distribution is either flatter or more peaked than normal; the *kurtosis* of the normal distribution is 0. From Table 2 we see that since both skewness and kurtosis are significantly different from zero (47.95/3354.54) the fact that LoS dataset is clearly non-normally distributed is once again proved.

Next, the results obtained by applying the agglomerative hierarchical clustering methods are presented.

The key operation is the computation of the proximity of two clusters, and different approaches to defining the distance between clusters distinguish the different algorithms. In this cluster analysis we have used the *single linkage* / Euclidian distance (the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters).

Figure 3 and 4 below depict the dendrogram corresponding to different cut-off distances.

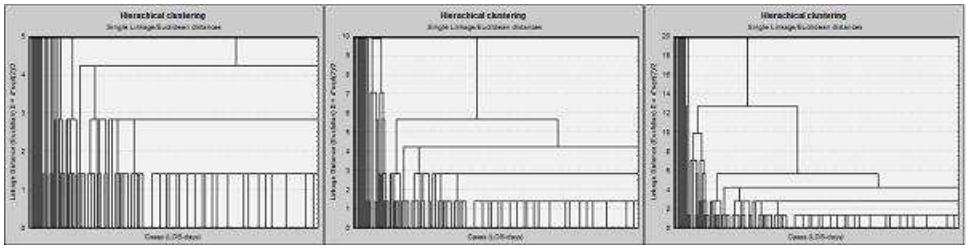

Figure 3. Agglomerative hierarchical clustering (distance from 5 to 20)

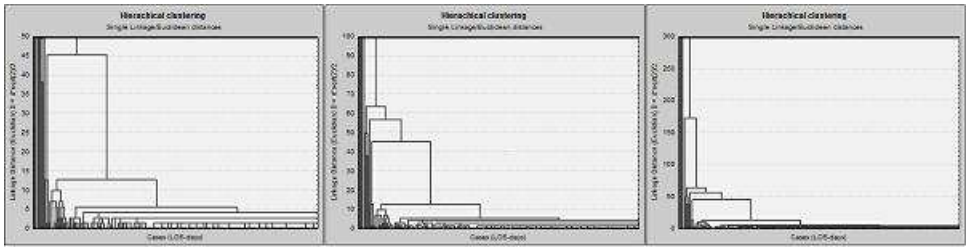

Figure 4. Agglomerative hierarchical clustering (distance from 50 to 300)

It is easy to see from the two figures that the majority of data (days spent in hospital) lies between zero and ten days (result also obtained by a simple statistical summary). Larger the number of days spent in hospital, smaller number of clusters. Plotting the graph of amalgamation schedule we obtain useful information concerning the way the agglomeration of data take place. Thus, the graph of amalgamation

schedule displays a line graph of the linkage distances across consecutive steps of the linking process. This graph is useful for identifying plateaus where many clusters are formed at approximately the same linkage distance - our case. This may indicate a natural "discontinuity" in terms of distances between the observed objects. Figure 5 displays such a behavior.

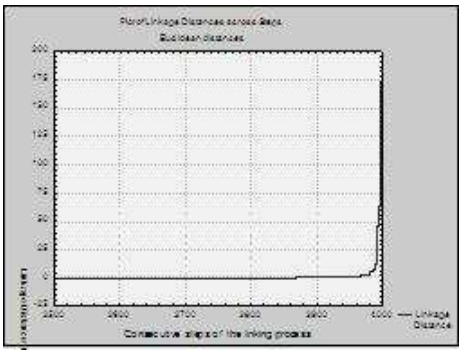

Figure 5. Graph of amalgamation schedule

It is easy to see from this graph that for a large number of steps of the linking process the corresponding linkage distance remains very small, indicating very large data with similar size (LoS between zero and 10-15 days). After that, for few steps, the corresponding distance grows very fast, indicating the outliers in data.

5. Conclusions

This paper introduced the problem of grouping patients spells according to their LoS using the agglomerative hierarchical clustering algorithm. More research is needed for deciding whether other clustering algorithms give better results when applied on this database.

References

[1] **Belciug S.**, Patients Length of Stay Grouping using K-means algorithm, First Doctoral Student Workshop, Pitesti, May 2009

[2] **Bagust, A., Place, M. and Posnett, J.W.** ,Dynamics of bed use in accommodating emergency admissions: Stochastic simulation model, British Medical Journal 319, 155-158, 1999.

[3] **Horrocks, P.**, The components of a comprehensive district health service for elderly people - a personal view, Age and Ageing 15, 321-342, 1986.

[4] **McQuarrie, D.G.**, Hospitalization utilization level: The application of queuing theory to a controversial medical economic problem, Minesota Medicine 66, 679-686, 1983.

[5] **Vasilakis C. & Marshall A.H.**Modelling nationwide hospital length of stay: opening the black box. The Journal of the Operational Research Society, 56, 862-869, 2005.

(Smaranda Belciug) Faculty of Mathematics and Computer Science, University of Craiova, Romania
E-mail address: smaranda.belciug@inf.ucv.ro