Annals of University of Craiova, Math. Comp. Sci. Ser.
Volume 36(2), 2009, Pages 19–24
ISSN: 1223-6934

Evaluation on liver fibrosis stages using the k-means clustering algorithm

Marina Gorunescu, Florin Gorunescu, Radu Badea, and Monica Lupsor

Abstract. Hepatic fibrosis, seen as an indicator in the progression of chronic liver disease, is measured on a five level scale, using the direct biopsy. The goal of this paper is to develop a k-means clustering methodology used for patient segmentation in accordance with the fibrosis levels. The study aims to assess the effectiveness of such a patient clustering, based on the main biochemical parameters and stiffness values, compared with the standard rule given by biopsy.

1. Introduction

Hepatic fibrosis is an indicator in the progression of chronic liver disease. Fibrosis itself causes no symptoms but can lead to portal hypertension -the scarring distorts blood flow through the liver- or cirrhosis -the failure to properly replace destroyed liver cells results in liver dysfunction. Thus, the last stage of liver stiffness leads to cirrhosis and hepatocellular carcinoma. Diagnosis is based on liver biopsy and treatment involves correcting the underlying condition when possible. Liver biopsy is currently the only means of detecting hepatic fibrosis, being indicated to clarify the diagnosis and to stage its progress (e.g. in chronic hepatitis C, whether fibrosis has progressed to cirrhosis). Noninvasive tests (e.g. serologic markers) are under study but are not yet ready for routine clinical use. Imaging tests such as ultrasonography, CT, and MRI may detect findings associated with fibrosis [5]. One of the last technological discovery worldwide in the evaluation of hepatic fibrosis is the Fibroscan (Echosens, Paris, France), a specially adapted ultrasound device using the principle of the one-dimension transient elastography (TE) for the assessment of liver stiffness. Existing patient grouping systems have been developed using either clinical opinion and/or statistical analysis. Although it may be highly desirable to automate the process of deriving statistically valid and clinically meaningful patient grouping systems, it must be taken into account that groups based solely on statistical analysis often result in groups which do not necessarily make sense clinically [1]. As such, it has been recognized that in order to develop a practical grouping methodology, a combination of clinical input and robust statistical methods is required [4].
As patient grouping techniques, clustering algorithms have been used in the context of health care to better understand the relationships between data when the groups are neither known nor cannot be predefined. These algorithms essentially derive data of similar type based on some measure of alikeness or closeness. Examples of clustering algorithms include hierarchical methods such as BIRCH [7], density-based methods such as DBSCAN [2], model-based methods such as mixture density modeling [6], and partition methods such as the k-means algorithm [3].
In this paper we use the k-means algorithm to cluster patients in five groups, based on

the main biological parameters and stiffness values obtained by Fibroscan, in accordance with the five classes of liver fibrosis given by direct biopsy. Thus, we compare the patient segmentation obtained by biopsy with that obtained by the k-means algorithm using the standard medical parameters. Valuable conclusions were drawn from this comparison of significantly different methodologies.
The rest of the paper is organized as follows. The next section summarizes the main characteristics of the k-means clustering algorithm. In Section three we present a real-life database, used to assess our findings. Section four discusses the results and finally we report our conclusions in section five.

2. k-means clustering algorithm

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 certain 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. The algorithm aims to minimise an objective function, given by the squared error function:

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

where the centroid c_j is defined by:

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

where $|| \cdot ||$ is a 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 algorithm stops when the centroids remain unchanged between two consecutive iterations or the squared error does not improve significantly.

3. Dataset

The dataset used in this study consists of 743 consecutive patients with chronic HCV infection, examined at the 3rd Medical Clinic, University of Medicine and Pharmacy Cluj-Napoca, Romania, between May 2007 and August 2008. All of them have positive HCV-RNA in their serum and underwent percutaneous liver biopsy for grading and staging the diseases. Moreover, all patients were referred to liver stiffness measurement. Besides the epidemiological, anthropometric clinical parameters and other important predictive parameters, the following biological parameters were determined for all patients on the same day as liver stiffness (i.e. the Fibroscan output): stiffness, sex, age, body mass index (BMI), glycemia, triglycerides, cholesterol, aspartate aminotransferase (AST), alanin aminotransferase (ALT), gamma glutamyl transpeptidase (GCT), total bilirubin (TB), alkaline phosphatase (AP), prothrombin index (PI), tocopheryl quinines (TQS), prothrombin time ratio/international normalized (INR), prolonged activated partial thromboplastin time, haematids (erytthrocytes), hemoglobin, hematocrit, medium erytthrocity volume, average eritrocitary hemoglobin, average concentration of hemoglobin in a red blood cell, thrombocytes, sideraemia, high density lipoprotein cholesterol. These parameters represent the predictive factors used as attributes in the clustering process. The liver biopsy output,

represented by the five Metavir F values (MF0 to MF4), was considered as comparison factor to the k-means clustering segmentation, with k equaling 5, that is the number of Metavir F values. The aim of this study was to assess the effectiveness of an automatic segmentation (k-means clustering), based on the above parameters, in comparison to the 'standard gold' methodology, based on the direct biopsy.
The study was approved by the local Ethical Committee of the University of Medicine and Pharmacy "Iuliu Hatieganu" Cluj-Napoca. The nature of this study was explained to the patients, each of whom provided written informed consent before the beginning of the study in accordance with the principles of the Declaration Helsinki (revision of Edinburgh, 2000).

4. Results

We have performed the k-means clustering algorithm using the above 25 attributes. The clustering result is displayed in Table 1. The rows here represent the actual classes obtained by direct biopsy (MF0-MF4), and the columns represent the predicted clusters (C1-C5), obtained based on the k-means algorithm. Each cell contains both the number of patients from the actual class belonging to the predicted cluster, together with the corresponding percentage. In this way, more detailed information on misclassification is provided.

Table 1. Confusion (misclassification) table

	C1	C2	C3	C4	C5
MF0(29)	1 (3.45%)	4 (13.80%)	24 (82.76%)	0 (0%)	0 (0%)
MF1(234)	45 (19.23%)	14 (5.98%)	175 (74.79%)	0 (0%)	0 (0%)
MF2(175)	40 (22.86%)	16 (9.14%)	117 (66.86%)	1 (0.57%)	1 (0.57%)
MF3(88)	11 (12.50%)	16 (18.18%)	59 (67.05%)	2 (2.27%)	0 (0%)
MF4(217)	25 (11.52%)	109 (50.23%)	71 (32.72%)	12 (5.53%)	0 (0%)

From Table 1 above we can see that the large majority of patients are distributed in the cluster C3 (83 % with MF0, 75 % with MF1, 67 % with MF2 and MF3, and just 32 % with MF4), only the cluster C2 containing half of MF4 patients. Let us also note the closeness of cluster C1 to the main cluster C3. This means that, although they have different MF scores, the diagnosis (patients segmentation) based on the closeness of their biological measures is not enough to provide a good accuracy.

5. Analysis of variance.

The goal of the k-means clustering procedure is to classify objects into a user-specified number of clusters. To evaluate the appropriateness of the classification, we can compare the within-cluster variability -small if the classification is good - to the between-cluster variability - large if the classification is good. Using the analysis of variances (F-ratios), we obtained that all the 24 explanatory factors excepting total bilirubin (P-level = 0.17) are significant (P-level < 0.02) for a good segmentation of patients, under these circumstances.

6. Graph of means.

The graph of means displays the line graph of the means across clusters. This graphical technique is very useful for visually summarizing the differences in means between clusters.

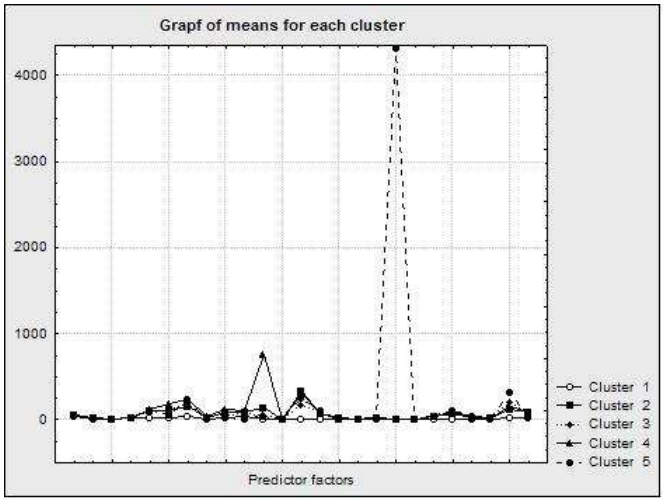

Figure 1. Graph of means for the five clusters

From the graph of means we can visually conclude that clusters C1, C2 and C3 are very close, emphasizing the observation from Table 1 regarding the distribution of patients among the five clusters.
The average stiffness (Fibroscan output) of patients in each cluster is given in Table 2 below.

Table 2. Clustering vs. average stiffness

Clustering vs. average stiffness					
	C1	C2	C3	C4	C5
Stiffness	12.07	28.25	9.7	30.43	4.2

Table 2 shows that the large majority of patients, irrespective their MF stage, belonging to cluster C3, have a mean stiffness equaling 9.7. In this case, the closest cluster is C1, in accordance with Table 1 and Figure 1. Cluster C2 is represented by an average stiffness equaling 28.25, far from those of clusters C1 and C3. Observe that, different from Figure 1, which displays the graph of means depending on each of 25 predicting attributes (including the stiffness), Table 2 shows the average stiffness in each cluster only. Thus, in this table, the relation between the clustering process and the Fibroscan output (averaged) is shown, emphasizing the connection between the Fibroscan technique and the patient segmentation
Finally, in Table 3 we summarized the Euclidean distances between cluster centers.

Clustering vs. average stiffness					
	C1	C2	C3	C4	C5
C1	0	87.19	69.51	176.46	4.2
C2	87.19	0	39.97	127.73	863.31
C3	69.51	39.97	0	144.81	862.86
C4	176.46	127.73	144.81	0	875.90
C5	867.66	863.31	862.86	875.90	0

It is easy to see from the above table that clusters C1, C2 and C3 are close enough, based on the distance between centroids. The closest centroids are those of C2 and C3, followed by those of C1 and C3. These results are concordant with the observations from Tables 1 and 2, and Figure 1. Thus, these three clusters are similar enough, although they contain patients from all the MF stages.

7. Conclusions

This paper introduced the problem of grouping patient according to their main biological parameters and the Fibroscan output, using the k-means clustering algorithm, and analyzed the relation between such segmentation and the classification given by liver biopsy. Based on the results in this paper, the k-means clustering algorithm appears to be a viable approach for grouping patients according to standard medical parameters, providing valuable knowledge about the relation between the liver fibrosis and medical features. On the other hand, clustering patients based only on the above medical parameters is not enough to obtain good accuracy for the fibrosis level. Thus, other parameters have to be considered to obtain concordance between the biopsy result and the automatic patient grouping.

8. Acknowledge.

This paper was supported through the research grant no. 41071/2007, entitled *"Predictive diagnosis algorithm for the hepatic fibrosis stages evolution using non-invasive ultrasonographic techniques, optimized by stochastic analysis and imaging techniques - SONOFIBROCAST"*, financed by the National Center of Programs Management (CNMP), Ministry of Education, Research and Innovation, Romania.

References

[1] **Averill, R.F.**, DRGs: Their Design and Development, Health Administration Press, 1991.
[2] **Ester, M., Kriegel, H., Sander, J., Xu, X.**, A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd International Conference on Knowledge Discovery and Data Mining, 226-231, 1996.
[3] **Han, J., Kamber, M.**, Data Mining: Concepts and techniques. Morgan Kaufmann, San Francisco, USA, 2006.
[4] **Kulinskaya, E.**, International Casemix Research: Why and How. Proc. 19th International Case Mix Conference, Washington, DC, 2003.
[5] Merck Manual of Diagnosis and Therapy, Online version: `http://www.merck.com/mmpe/sec03/ch026/ch026b.html`, 2007.
[6] **McLachlan, G.J, Peel, D.**, Finite Mixture Models, John Wiley & Sons, New York, 2000.
[7] **Zhang, T., Ramakrishnan, R., Livny, M.**, BIRCH: An efficient data clustering method for very large databases. SIGMOD '96, Proc. 1996 ACM SIGMOD International Conference on Management of Data, New York, USA. 103-114, 1996.

(Marina Gorunescu) Faculty of Mathematics and Computer Science, University of Craiova
E-mail address: mgorun@inf.ucv.ro

(Florin Gorunescur) Department of Biostatistics and Computer Science, University of Medicine and Pharmacy of Craiova
E-mail address: gorun@umfcv.ro

(Radu Badea) Department of Hepatology, 3rd Medical Clinic, University of Medicine and Pharmacy "Iuliu Hatieganu" Cluj-Napoca
E-mail address: rbadea@umfcluj.ro

(Monica Lupsor) Department of Hepatology, 3rd Medical Clinic, University of Medicine and Pharmacy "Iuliu Hatieganu" Cluj-Napoca
E-mail address: mmlupsor@yahoo.com