ISSN:0975-1459
Journal of Pharmaceutical Sciences and Research
www.jpsr.pharmainfo.in

Machine Learning in Pharmaceutical Research: Data Clustering, Why so and How so

Ton J Cleophas, MD, PhD,
Professor ,*European College Pharmaceutical Medicine,*
Lyon France

Abstract

Background:
In clinical data subgroups can sometimes be identified using regression analysis of subgroup characteristics against some outcome variable, but in data samples without an available outcome variable cluster analysis is a suitable alternative. It is based on the concept that patients with closely related characteristics may also be more related in other fields like prognoses and treatment efficacies.

Objective:
To compare the performance of three different cluster methodologies, hierarchical , k-means, and density-based clustering.

Methods:
A simulated data example of fifty patients with mental depression was used.

Results:
Each cluster methodology identified three clusters. However, the cluster patterns were very different. The hierarchical method showed round patterns different in size, the k-means method round patterns equal in size, and the density-based method non-circular patterns also different in size. The patterns from the hierarchical method were better in agreement with the patterns as clinically expected, than those from the other methods.

Conclusions:
1. Cluster analysis is little used in clinical research.
2. Hierarchical cluster is adequate if subgroups in the data are expected to be different in size but, otherwise, Gaussian-like. It is available in the module Classify of SPSS.
3. K-means cluster analysis is adequate if subgroups are expected to be approximately similar in size. It is also available in the module Classify of SPSS.
4. Density-based cluster analysis is adequate if small outlier groups between an, otherwise, homogeneous population is expected. It is not available in SPSS, but an interactive JAVA Applet is freely obtainable at the Internet.

Keywords: cardiovascular research, machine learning, hierarchical clustering, k-means clustering, density-based clustering

INTRODUCTION

Populations with a single clinical diagnosis are, otherwise, often very heterogeneous. This may have consequences regarding prognosis and treatment efficacies. E.g., patients with mental depression may suffer from subtypes like reactive depression, depression with insomnia or true depression.[1] Patients with gastric cancer may have different expression levels of genes that are related with their prognoses.[2] Different characteristics in a population of HIV patients were held responsible for their HIV vaccine efficacy.[3] Underlying mechanisms were established for explaining high anti-trypanosomal drug efficacy.[4] Many more examples can be given. Subgroups can sometimes be identified using regression analysis of potential subgroup properties against some outcome variable, but in data samples without an available outcome variable cluster analysis is a suitable alternative. It is based on the concept that patients with closely related characteristics might also be more related in other fields like prognoses and drug efficacies.

Unlike regression analysis, cluster analysis does not require a dependent (outcome) variable. In a sense the patients themselves are the dependent variables. Cluster analysis is currently an important methodology in explorative data mining, and a main task in machine learning, and is sometimes called unsupervised machine learning, because there is, generally, no dependent variable.[5] It is widely used by econometrists and sociologists for identifying population subgroups[6], but in clinical research it is virtually unused. Apart from its current key role in sequence-clustering[7,8], which is a method for clustering related DNA and protein sequences, we found only sporadic published papers of cluster analysis in any type of health research. Two adverse event studies[9,10], one drug manufacturing study[11], and one patient compliance study[12] have been published.

The current paper using a simulated example assesses the potential of cluster analysis for the analysis of clinical data and compares the clustering results of different methodologies, including hierarchical, k-means, and density-based clustering.

Three clustering methodologies

Three methodologies are currently used.

1. Hierarchical cluster analysis

It was invented by Robert Sibson (1973), statistician from King's College Cambridge UK statistical department[13] and Daniel Defays (1977), psychologist from Liege University Belgium.[14] A cluster is estimated by the distances between the values needed to connect the cases. The smaller the distance, the more similar the cases are. The distance is calculated as the squared difference between two cases. The method starts with all patients being a cluster of his/her own. Then, the smallest distances are used to form the first clusters.

This procedure continues, and stops when all patients are in a cluster. With Gaussian-like data as commonly observed in scientific research, the clusters tend to have an oval pattern and with similarly sized scales even a round pattern, but they are not equal in size.

2. K-means cluster analysis

It was invented by Stuart Lloyd, a physicist from New Jersey who worked at the Math Department of Bell

Telephone in 1957, but was first published in 1982.[15] Compared to hierarchical clustering this method works in the opposite direction, but, otherwise, largely similar. It does not start with all patients being a cluster of his/her own, but instead, randomly selects cluster centers, and, then by iteration tries and finds the best fit centers for the data given, i.e., those with the shortest distances to the centers. Intuitively one may assume that this procedure should lead to the same result. However, this is not necessarily true. The point is that one important assumption of the k-means method is that the clusters are equally sized, and this is not an assumption of the hierarchical method.

3. Density-based cluster analysis

It was invented by Martin Ester and Hans Kriegel, professors of computer science at Muenich University in 1996.[16] Density-based clusters are defined as areas of higher density than the remainder of the data. Individuals in the sparse area are considered as noise (random effects). Density-based clustering connects points that satisfy a density criterion given by a minimum number of patients within a defined radius. Unlike in the above two methods, here the clusters do not need to be round, but they are multiform areas that are, simply, more dense than the cluster-less areas.

Example

Fifty patients with mental depression are assessed for age and depression score (zero = very mild, 10 = severest). We will use various cluster methods in order to identify clusters with different ages and severities. We have some prior idea about differences in age and severity between patients with true depression, reactive depression, and depression with insomnia. Table 1 gives the patient data.

Table 1. Data file of the example used, patients are used cases, cluster membership of the hierarchical and k-means clustering methods.

Age	*Depression Score*	*patient number*	*Hierarchical clustering*	*k-means clustering*
20,00	8,00	1	1	1
21,00	7,00	2	1	1
23,00	9,00	3	1	1
24,00	10,00	4	1	1
25,00	8,00	5	1	1
26,00	9,00	6	1	1
27,00	7,00	7	1	1
28,00	8,00	8	1	1
24,00	9,00	9	1	1
32,00	9,00	10	1	1
30,00	1,00	11	1	1
40,00	2,00	12	2	2
50,00	3,00	13	2	2
60,00	1,00	14	3	2
70,00	2,00	15	3	3
76,00	3,00	16	3	3
65,00	2,00	17	3	3
54,00	3,00	18	3	2
54,00	4,00	19	3	2
49,00	3,00	20	2	2
30,00	4,00	21	1	1
25,00	5,00	22	1	1
24,00	4,00	23	1	1
27,00	5,00	24	1	1
35,00	6,00	25	1	1
45,00	5,00	26	2	2
45,00	6,00	27	2	2
67,00	7,00	28	3	3
80,00	6,00	29	3	3
80,00	5,00	30	3	3
40,00	1,00	31	2	2
50,00	2,00	33	3	2
80,00	4,00	34	3	3
50,00	5,00	35	2	2
76,00	6,00	36	3	3
65,00	7,00	37	3	3
79,00	8,00	38	3	3
57,00	3,00	39	3	2
46,00	4,00	40	2	2
54,00	5,00	41	3	2
74,00	6,00	42	3	3
65,00	7,00	43	3	3
57,00	9,00	44	3	2
68,00	8,00	45	3	3
67,00	7,00	46	3	3
65,00	6,00	47	3	3
64,00	5,00	48	3	3
74,00	4,00	49	3	3
75,00	3,00	50	3	3

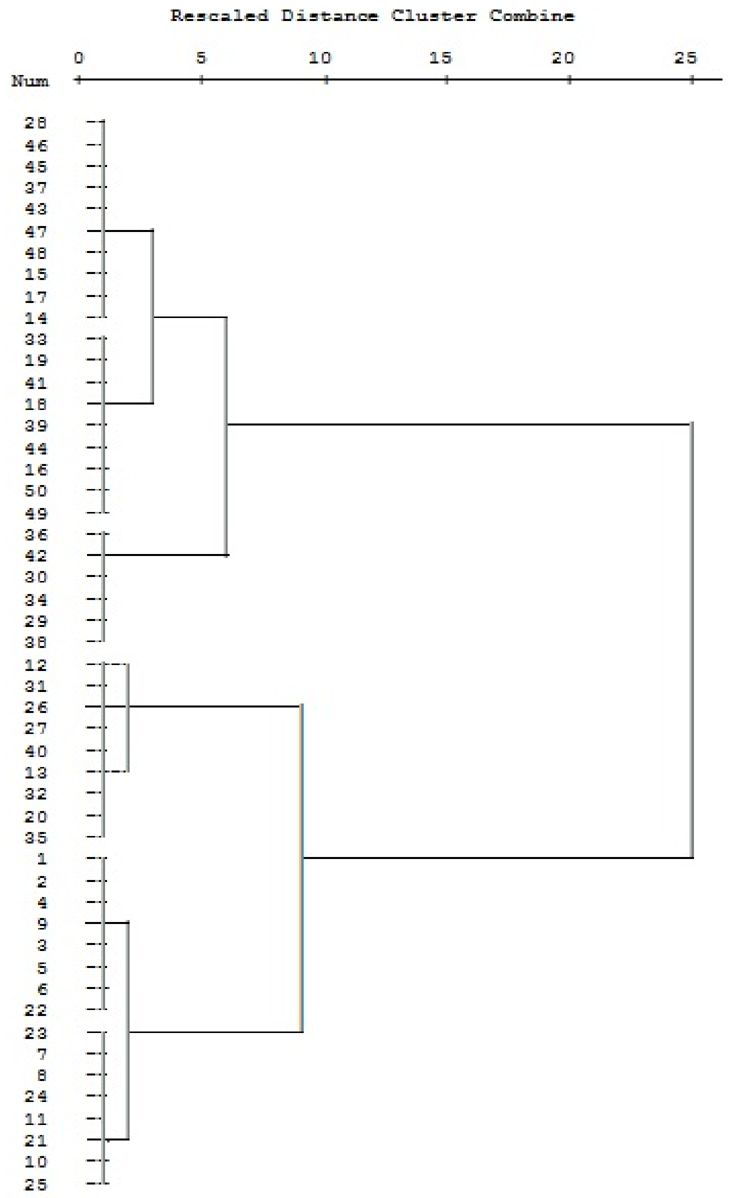

Figure 1. Dendrogram of the 50 cases of Table 1. The actual distances between the cases are rescaled to fall into a range of 0 to 25 (0 = minimal distance, 25 = maximal distance). The cases 1-11, 21-25 are clustered together in cluster 1, the cases 12, 13, 20, 26, 27, 31, 32, 35, 40 in cluster 2, both at a rescaled distance from 0 at 3. The remainder of the cases are clustered at a distance of 6. At that point, three clusters of cases have been indentified with cases more similar to one another than to the cases of the other clusters. Beyond the distance of 10 only two clusters are left in the data.

Table 2. The three clusters identified by the k-means cluster model were very significantly different from one another both by testing the y-axis (depression score) and the x-axis variable (age).

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Age	8712,723	2	31,082	47	280,310	,000
Depression Score	39,102	2	4,593	47	8,513	,001

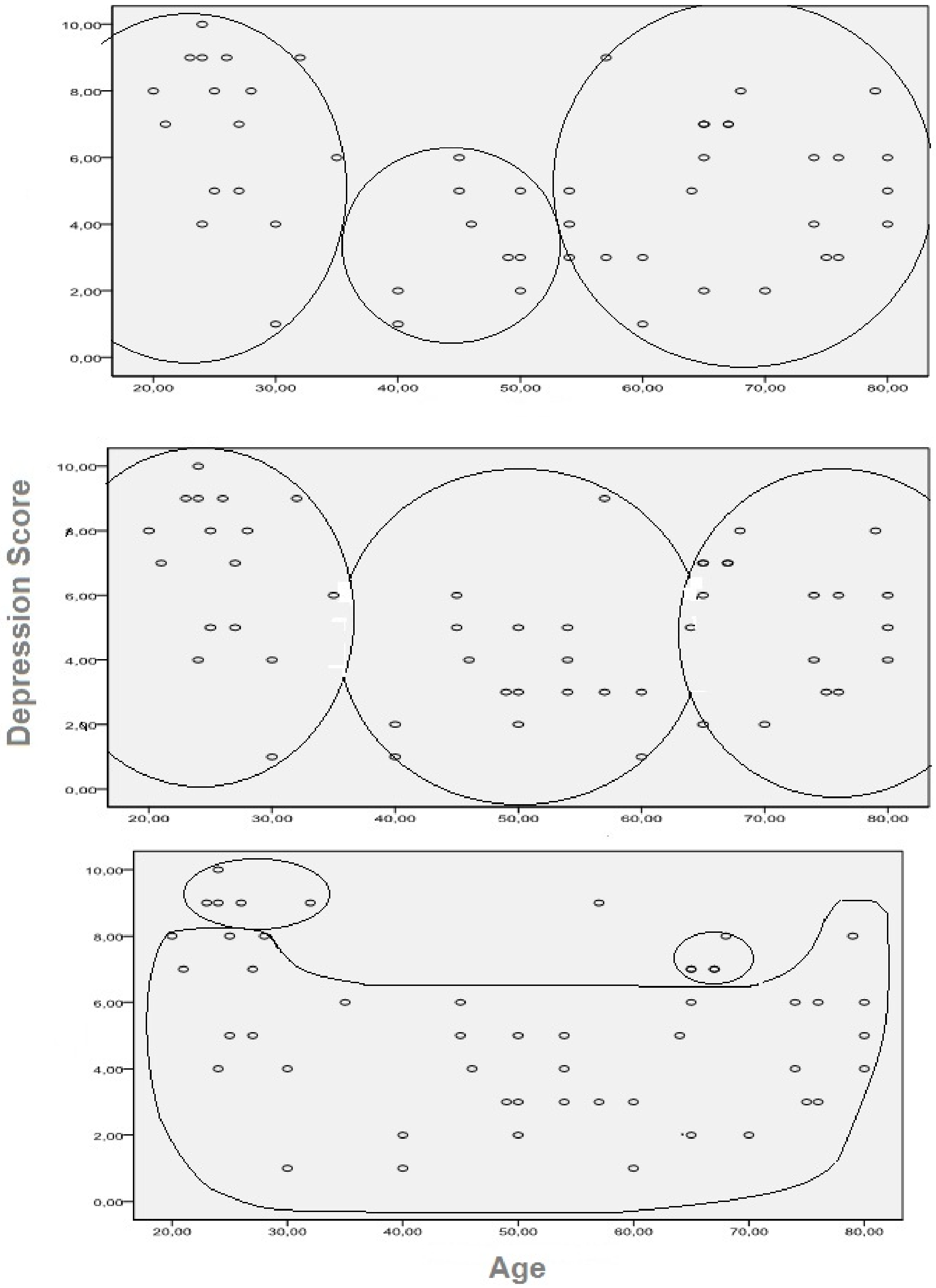

Figure 2. Graphs of the data from the example in this paper: upper graph hierarchical cluster analysis, middle graph k-means cluster analysis, lower graph density-based cluster analysis.

Hierarchical cluster analysis

Patients are called cases. The distances between the cases are calculated as the squared differences between two cases. We will use SPSS statistical software.[17]

Command: Analyze….Classify….Hierarchical Cluster Analysis….enter variables….Label Case by: case variable with the values 1-50….Plots: mark Dendrogram….Method….Cluster Method: Between-group linkage….Measure: Squared Euclidean Distance….OK.

Figure 1 shows a dendrogram from the data from Table 1. The actual distances between the cases are rescaled to fall into a range of 0 to 25 units (0 = minimal distance, 25 = maximal distance). The cases no. 1-11, 21-25 are clustered together in cluster 1, the cases 12, 13, 20, 26, 27, 31, 32, 35, 40 in cluster 2, both at a rescaled distance from 0 units at approximately 3 units (Figure 2). The remainder of the cases are clustered at approximately 6 units. Obviously, three clusters of cases have been indentified with cases more similar to one another than to the cases of the other clusters. When minimizing the output file, the data file comes up and it now shows the cluster membership of each case. We use SPSS to draw a Dotter graph of the data.

Command: Analyze….Graphs….Legacy Dialogs: click Simple Scatter….Define….Y-axis: enter Depression Score….X-axis: enter Age….OK.

The graph produced by SPSS is given in triplicate in Figure 1, and the memberships of the cases per method is shown. The upper graph, the hierarchical model, shows that all of the clusters are oval and even approximately round because variables have similarly sized scales, but they are different in size. Two large clusters are in both the youngsters and the elderly, one small cluster is in between. This hierarchical cluster model is in agreement with the patterns as clinically expected: two large populations with respectively true depression (younger patients) and reactive depression (elderly), and one small population with depression associated with insomnia. The method does not provide a statistic to test whether the three clusters are significantly different from one another, but the graph shows that there is a complete separation between the three clusters, and, so, the between-cluster differences must be statistically very significant. No statistical test is needed.

K-means cluster analysis

Compared to hierarchical clustering this method works in the opposite direction. It does not start with all patients being a cluster of his/her own, but instead, randomly selects cluster centers, and, then by iteration tries and finds the best fit centers for the data given. Intuitively one may assume that this procedure should lead to the same result. However, this is not necessarily true. The point is that one important assumption of the k-means method is that the clusters are equally sized, and this is not an assumption of the hierarchical method. SPSS is used again for analysis.

Command: Analyze….Classify….K-means Cluster Analysis….Variables: enter Age and Depression score….Label Cases by: patient number as a string variable….Number of clusters: 3 (in our example chosen for comparison with the above method)….click Method: mark Iterate….click Iterate: Maximal Iterations: mark 10….Convergence criterion: mark 0….click Continue….click Save: mark Cluster Membership….click Continue….click Options: mark Initiate cluster centers….mark ANOVA table….mark Cluster information for each case….click Continue….OK.

Table 2 shows that the three clusters identified are very significantly different from one another, both by testing the y (depression score) and the x variables (age) against the cluster membership. When minimizing the output file the data file comes up, and it now shows the cluster membership of each case 1-50. It can be observed that there is a lot of agreement between the memberships of the hierarchical and k-means methods, but there are differences, particularly in the patients between 50 and 60 years of age: they were fully assigned to a different cluster. The middle graph of Figure 2 shows what happened. The left part of the elderly population is now assigned to the insomnia population. Also, the three clusters are now equal in size. Indeed the k-means procedure assumes equal sizes, and the result shows that this assumption is satisfied by the analysis. But why should clusters of a random sample of patients with true depression, insomnia, and reactive depression be equal in size. The best way to find out would be to repeat the study using a larger random sample, but this is laborious and costly. The next best solution has already been performed, and is, actually, hierarchical cluster analysis, because it only uses the neighbourhood criterion and skips the equal size criterion.

Density-based clustering

The DBSCAN method was used (density based spatial clustering of application with noise).[17]

As this method is not available in SPSS, an interactive JAVA Applet freely available at the Internet was used.[18] The DBSCAN connects points that satisfy a density criterion given by a minimum number of patients within a radius given (radius = Eps; minimum number = Min pts).

Command: User Define….Choose data set: remove values given….enter you own x and y values….Choose algorithm: select DBSCAN….Eps: mark 25….Min pts: mark 3….Start….Show.

Three clusters are shown (Figure 1 bottom graph). Two very small ones, one with very high depression scores in youngsters and one with very high depression scores in patients 60-70 years are observed, and one large one with moderate to low levels of depression at all ages. All of the clusters identified are non-circular and, are, obviously, based on differences in patient-density.

DISCUSSION

Cluster analysis is, currently, an important methodology in explorative data mining, and a main task in machine learning[4-6], but, unfortunately, little used in health research, in spite of the omnipresence of heterogeneities in patient diagnosis groups. The little use in pharmaceutical research is probably due to the traditional belief of pharmaceutical investigators in clinical trials, where randomization takes care that heterogeneities in the data are equally distributed between the treatment and control groups, and where they are no further taken into account. Controlled clinical trials may, indeed, be more accurate and reliable for making health predictions, but are uncontrolled data completely meaningless? Even if heterogeneities established are clinically relevant in less than 10% of the cases, 10 % is better than 0%. Also, 10% is a lot, if you consider the ready availability of large and complex data files in electronic health records of modern health facilities and other institutions.

In the example of this paper we have argued that hierarchical clustering was the best way for assessing the data, because of a prior belief, that the clusters were likely to be different in size, and, because we had no arguments for non-Gaussian patterns in these data.

However, the other two methods may be more appropriate with other types of data. For example, the k-means method might be more appropriate with clusters having the same size like clusters of genders in a random population. Density-based clustering may be more appropriate with large homogeneous populations and one or more relatively small outlier subsets, like patients with specific environmental, genetic, life style characteristics etc.

Only two-dimensional clusters are reviewed here, with age and depression severity as the only variables. However, for all of the three models reviewed in this paper multidimensional clustering is possible, if you wished to include more than two variables. Multidimensional clustering is relevant in clinical research considering the multifactorial nature of disease and drug efficacy[1-6], and can be performed even if outcome variables are not available.

We should add that two-dimensional density-based clustering[19] and, in some studies, also two-dimensional k-means clustering[20] were of great importance in the field of imaging, like image compression and image color quantization.[6] Unfortunately, it is little used in medical imaging like PET (positive emission tomography) and MRI (magnetic resonance image) scanning[6], but this is a matter of time, now that it is increasingly available in SPSS and other statistical software programs.

We conclude.

1. Cluster analysis is little used in clinical research.
2. Hierarchical cluster is adequate if subgroups in the data are expected to be different in size but, otherwise, Gaussian-like. It is available in the module Classify of SPSS.
3. K-means cluster analysis is adequate if subgroups are expected to be approximately similar in size. It is also available in the module Classify of SPSS.
4. Density-based cluster analysis is adequate if small outlier groups between an, otherwise, homogeneous population is expected. It is not available in SPSS, but an interactive JAVA Applet is freely available at the Internet.

REFERENCES

1. Anonymous. Diagnostic and statistical manual of mental disorders. Edited by the American Psychiatric Society, New York, 1978.
2. Solyanik GL. Multifactorial nature of tumor drug resistance. Exp Oncol 2010; 32: 181-5.
3. National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland. Enhancing HIV vaccine efficacy in high-risk drug users. Release data Jan 6 2003, RFA Number DA-03- 002.
4. Alsford S, Eckert S, Baker N, Glover L, Sanchez-Flores A, Leubg KF, Turner DJ, Field MC, Berriman M, Horn D. High throughput decoding of antitrypanosomal drug efficacy and resistance. Nature 2012; doi:10.1038/nature_10771.
5. Anonymous. Machine learning. http://en.wikipedia.org/Machine_learning, July 12-2012.
6. Anonymous. Cluster analysis. http://en.wikipedia.org/Cluster_analysis, July 12-2012.
7. Anonymous. Sequence clustering. http://en.wikipedia.org/wiki/Sequence_analysis, July 12 -2012.
8. Kim HK, Choi IJ, Kim HS, Oshima A, Michalowski A, Green JE. A gene expression signature of acquired chemoresistance to cisplatinum and fluorouracil combination chemotherapy in gastric cancer patients. Plos One 2011; 18: e16694.
9. Yeh ST. Clinical adverse event data analysis and visualization. Smith Kline Datafile, 10- 07-2012.
10. Bychowiec B, Piskorski J, Stanislawska K, Dziarmaga M, Mineczykowski A, Wykretowicz A, Wysocki H. An exploratory clustering study of rare adverse events in drug deluting stent patients. Comput Meth Science Technol 2010; 16: 5-11.

11 .Xu D, Redamn-Furey N. Statistical cluster analysis of pharmaceutical solvents. Int J Pharm 2007; 339: 175-88.

12 .Hawwa A, Millership JS, Collier PS, McCarthy A, Dempsey S, Cairns C, McElnay JC. Development of objective methodology to measure medication adherence to oral thiopurines in paediatric patients with acute lymphoblastic leucemia. Eur J Clin Pharm 2009; 65: 1105-12.

13. Sibson R. An optimally efficient algorithm for a single-link cluster method. Computer J 1973; 16: 30-4.
14. Defays D. An efficient algorithm for a complete link cluster method. Computer J 1977; 20: 364-6.

15 .Lloyd SP. Least square quantization in PCM. IEEE Transactions on Information Theory 1982; 28: 129-37.

16 .Ester M, Kriegel HP, Sander J, Xu X. A density based algorithm for discovering cluster in large spatial databases with noise. Proc 2nd Int Conf on Knowledge Discovery and Data Mining, Portland, Oregon, AAI Press, 1996.

17. SPSS statistical software. www.spss.com, 12-07-2012.
18. Data Clustering Applets. http://webdocs.cs.ualberts.ca/~yaling/Cluster/applet, 17-09-2012.
19. Anonymous. Density-based cluster and outlier analysis. www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering, 25-09,2012.

20 .Kanungo T, Mount D, Netanyahu N, Piatko C, Wu A. An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002; 24: 881-92.