Chapter 2
Density-Based Clustering to Identify Outlier Groups in Otherwise Homogeneous Data (50 Patients)

General Purpose

Clusters are subgroups in a survey estimated by the distances between the values needed to connect the patients, otherwise called cases. It is an important methodology in explorative data mining. Density-based clustering is used.

Specific Scientific Question

In a survey of patients with mental depression of different ages and depression scores, how does density-based clustering perform in identifying so far unobserved subgroups.

1	2	3
20.00	8.00	1
21.00	7.00	2
23.00	9.00	3
24.00	10.00	4
25.00	8.00	5
26.00	9.00	6
27.00	7.00	7
28.00	8.00	8
24.00	9.00	9
32.00	9.00	10
30.00	1.00	11
40.00	2.00	12
50.00	3.00	13
60.00	1.00	14
70.00	2.00	15
76.00	3.00	16

(continued)

T. J. Cleophas and A. H. Zwinderman, *Machine Learning in Medicine - Cookbook*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-319-04181-0_2,

(continued)

1	2	3
65.00	2.00	17
54.00	3.00	18

Var 1 age
Var 2 depression score (0 = very mild, 10 = severest)
Var 3 patient number (called cases here)
Only the first 18 patients are given, the entire data file is entitled "hierk-meansdensity" and is in extras.springer.com.

Density-Based Cluster Analysis

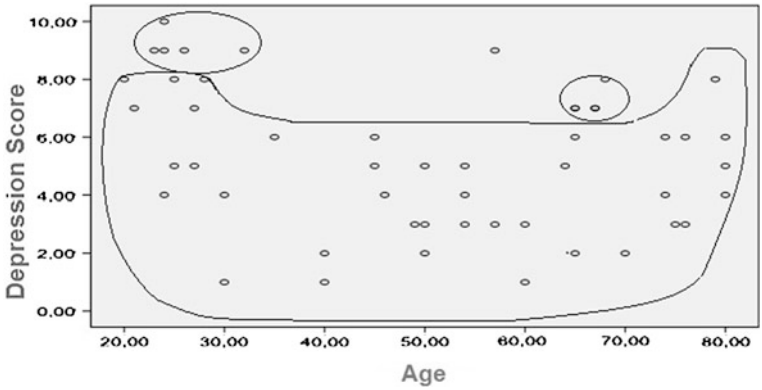
The DBSCAN method was used (density based spatial clustering of application with noise). As this method is not available in SPSS, an interactive JAVA Applet freely available at the Internet was used [Data Clustering Applets. http://webdocs.cs.ualberts.ca/~yaling/Cluster/applet]. The DBSCAN connects points that satisfy a density criterion given by a minimum number of patients within a defined radius (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 cluster memberships are again shown. We will use SPSS 19.0 again 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 (with age on the x-axis and severity score on the y-axis) shows the cases. Using Microsoft's drawing commands we can encircle the clusters as identified. Two very small ones, one large one. All of the clusters identified are non-circular and, are, obviously, based on differences in patient-density.



Conclusion

Clusters are estimated by the distances between the values needed to connect the cases. It is an important methodology in explorative data mining. Density-based clustering is suitable if small outlier groups between otherwise homogeneous populations are expected. Hierarchical and k-means clustering are more appropriate if subgroups have Gaussian-like patterns (Chap. 1).

Note

More background, theoretical and mathematical information of the three methods is given in Machine Learning in Medicine Part Two, Chap. 8. Two-dimensional Clustering, pp. 65–75, Springer Heidelberg Germany 2013. Hierarchical and k-means clustering are reviewed in the previous chapter.

http://www.springer.com/978-3-319-04180-3

Machine Learning in Medicine - Cookbook

Cleophas, T.J.; Zwinderman, A.H.

2014, XI, 137 p. 14 illus., Softcover
ISBN: 978-3-319-04180-3