Chapter 3
Two Step Clustering to Identify Subgroups and Predict Subgroup Memberships in Individual Future Patients (120 Patients)

General Purpose

To assess whether two step clustering of survey data can be trained to identify subgroups and subgroup membership.

Specific Scientific Question

In patients with mental depression, can the item scores of depression severity be used to classify subgroups and to predict subgroup membership of future patients.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9
9,00	9,00	9,00	2,00	2,00	2,00	2,00	2,00	2,00
8,00	8,00	6,00	3,00	3,00	3,00	3,00	3,00	3,00
7,00	7,00	7,00	4,00	4,00	4,00	4,00	4,00	4,00
4,00	9,00	9,00	2,00	2,00	6,00	2,00	2,00	2,00
8,00	8,00	8,00	3,00	3,00	3,00	3,00	3,00	3,00
7,00	7,00	7,00	4,00	4,00	4,00	4,00	4,00	4,00
9,00	5,00	9,00	9,00	2,00	2,00	2,00	2,00	2,00
8,00	8,00	8,00	3,00	3,00	3,00	3,00	3,00	3,00
7,00	7,00	7,00	4,00	6,00	4,00	4,00	4,00	4,00
9,00	9,00	9,00	2,00	2,00	2,00	2,00	2,00	2,00
4,00	4,00	4,00	9,00	9,00	9,00	3,00	3,00	3,00
3,00	3,00	3,00	8,00	8,00	8,00	4,00	4,00	4,00

Var 1–9 = depression score 1–9

This chapter was previously published in "Machine learning in medicine-cookbook 1" as Chap. 3, 2013.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_3

Only the first 12 patients are given, the entire data file is entitled “twostepclustering” and is in extras.springer.com.

The Computer Teaches Itself to Make Predictions

SPSS 19.0 is used for data analysis. It will use XML (eXtended Markup Language) files to store data. Now start by opening the data file.

Command:

Click Transform….click Random Number Generators….click Set Starting Point ….click Fixed Value (2000000)….click OK….click Analyze….Classify….TwoStep Cluster….Continuous Variables: enter depression 1-9….click Output: in Working Data File click Create cluster membership….in XML Files click Export final model….click Browse….File name: enter "export2step"….click Save….click Continue….click OK.

Returning to the data file we will observe that three subgroups have been identified and for each patient the subgroup membership is given as a novel variable, and the name of this novel variable is TSC (two step cluster). The saved XML file will now be used to compute the predicted subgroup membership in five future patients. For convenience the XML file is given in extras.springer.com.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9
4,00	5,00	3,00	4,00	6,00	9,00	8,00	7,00	6,00
2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00
5,00	4,00	6,00	7,00	6,00	5,00	3,00	4,00	5,00
9,00	8,00	7,00	6,00	5,00	4,00	3,00	2,00	2,00
7,00	7,00	7,00	3,00	3,00	3,00	9,00	9,00	9,00

Var 1–9 = Depression score 1–9

Enter the above data in a new SPSS data file.

Command:

Utilities….click Scoring Wizard….click Browse….click Select….Folder: enter the export2step.xml file….click Select….in Scoring Wizard click Next….click Use value substitution….click Next….click Finish.

The above data file now gives subgroup memberships of the five patients as computed by the two step cluster model with the help of the XML file.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10
4,00	5,00	3,00	4,00	6,00	9,00	8,00	7,00	6,00	2,00
2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00
5,00	4,00	6,00	7,00	6,00	5,00	3,00	4,00	5,00	3,00
9,00	8,00	7,00	6,00	5,00	4,00	3,00	2,00	2,00	1,00
7,00	7,00	7,00	3,00	3,00	3,00	9,00	9,00	9,00	2,00

Var 1–9 Depression score 1–9
Var 10 predicted value

Conclusion

Two step clustering can be readily trained to identify subgroups in patients with mental depression, and, with the help of an XML file, it can, subsequently, be used to identify subgroup memberships in individual future patients.

Note

More background, theoretical and mathematical information of two step and other methods of clustering is available in Machine learning in medicine part two, Chaps. 8 and 9, entitled “Two-dimensional clustering” and “Multidimensional clustering”, pp 65–75 and 77–91, Springer Heidelberg Germany 2013.