Clustering analysis of INDEX demographic database using AutoClass

Milan Nikolic (milan@cs.sfu.ca) and Dr. Ljiljana Trajkovic (ljilja@cs.sfu.ca)
Communication Networks Laboratory, School of Engineering Science, Simon Fraser University

AutoClass is an unsupervised Bayesian classification system that seeks a maximum posterior probability classification.
http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/

Input consists of a database of attribute vectors (cases) and a class model.

AutoClass finds the set of classes maximally probable with respect to the data and the model.

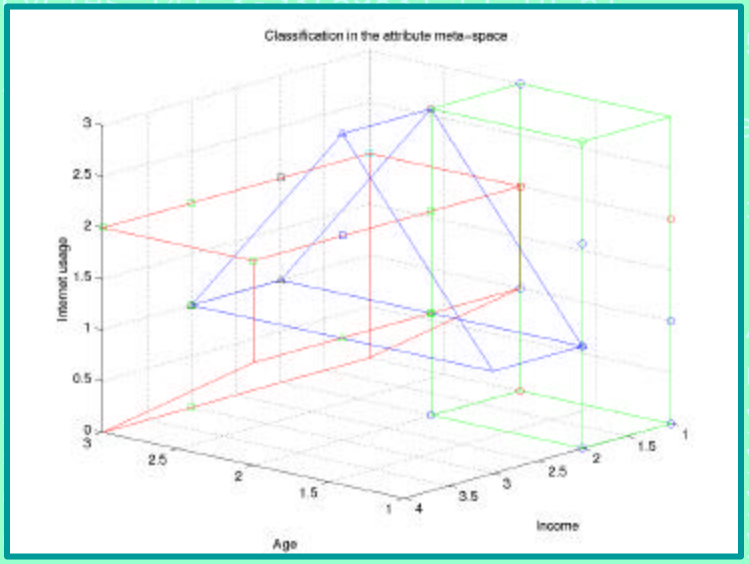
Output is a set of class descriptions and partial membership of the cases in the classes.

Age
1 : people in their 20's
2 : people in their 30's
3 : people in their 50's

Income
1 : low income
2 : mid income
3 : high income
4 : very high income

Internet usage
1 : employer pays for Internet usage at work
2 : university pays for Internet usage at work
3 : employer and university pay for Internet usage at work

Colours
green : professors
red : students
blue : technicians
black : administrators
cyan : none of the above



INDEX: Internet Demand Experiment is a market and technology trial at UC Berkeley that offers various qualities of service for Internet access.
http://www.index.berkeley.edu/public/index.phtml/

Analyzed database was collected from the INDEX project demographic questionnaire and consists of 84 cases with 23 attributes.

AutoClass found classification with 3 classes:

CLASS 1: ○, 47 cases
- students and technicians
- in their 20's or 30's
- with low or mid income

CLASS 2: □, 28 cases
- professors, technicians and administrators
- in their 30's or 50's
- with high or mid income
- Internet usage at home paid by university

CLASS 3: △, 9 cases
- employed people
- with high or mid income
- Internet usage at home paid by employer
- Internet usage at work paid by employer

Features:
- automatically determines the number of classes
- uses mixed discrete and real valued data
- handles missing values
- processing time is roughly linear in the amount of the data
- cases have probabilistic class membership
- allows correlation between attributes within a class
- predicts "test" case class membership from a "training" classification

References:
- R. Hanson, J. Stutz and P. Cheeseman, "Bayesian classification theory," Technical Report FIA-90-12-7-01, NASA Ames Research Center, Artificial Intelligence Branch, May 1991.
- P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): Theory and results," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., Menlo Park: The AAAI Press, 1995.
- J. Walrand and P. Varaiya, *High-Performance Communication Networks*, 2nd ed., San Francisco: Morgan-Kaufmann, 2000.
- R. Edell and P. Varaiya, "Providing Internet access: What we learn from INDEX," *IEEE Network*, vol. 13, no. 5, Oct. 1999, pp. 18-25.