December 2002

FUZZY CLUSTERING AS AN ALTERNATIVE CLASSIFICATION METHOD IN INFORMATION SYSTEMS RESEARCH

Vikram Sethi
University of Texas at Arlington

Mark Eakin
University of Texas at Arlington

Tonya Barrier
Southwest Missouri State University

Kevin Duffy
College of Business Administration

Vijay Sethi
Nanyang Technological University

FUZZY CLUSTERING AS AN ALTERNATIVE CLASSIFICATION METHOD IN INFORMATION SYSTEMS RESEARCH

Vikram Sethi
College of Business Administration
University of Texas at Arlington
vikram@@world.std.com

Mark Eakin
College of Business Administration
University of Texas at Arlington
eakin@uta.edu

Tonya Barrier
Southwest Missouri State University
tonyabarrier@smsu.edu

Kevin Duffy
College of Business Administration
kduffy@uta.edu

Vijay Sethi
Nanyang Business School
Nanyang Technological University
avsethi@ntu.edu

Abstract

Cluster Analysis is a powerful exploratory method that has been used in pattern analysis, grouping, decision-making, and pattern classification. Cluster analysis is particularly appropriate for the exploration of interrelationships among the data points to make an assessment of their structure. General tasks in cluster analysis (Jain, Murty, and Flynn 1999) involve pattern representation, definition of pattern proximity, clustering or grouping, and assessment of out. Within the broader scope of information systems research, cluster analysis has been used in the area of end-user computing, assessment of IS strategies, global information systems research and assessment of supply chain configurations. In most cases, researchers have adopted hard or crisp cluster analysis whereby each respondent or analysis unit is assigned a class label, or becomes a part of one and only one cluster (unitary membership). In several cases, unitary membership may not reflect situation realities. An alternative clustering methodology – fuzzy clustering - provide an alternative analytical mechanism to unitary membership by allowing gradual membership in different groups, with membership values indicating the probability or degree of membership within a specific group or cluster. We illustrate fuzzy clustering with an example from the field of end-user computing and then develop general guidelines for the broader use of the method.

Keywords: Cluster analysis, fuzzy clustering, classification methods, research methodologies

Cluster analysis is an important and powerful, mostly exploratory, tool to assess relationships among data points. The method involves the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. Generally, researchers have little prior knowledge regarding pattern structure, or they may have a specific typology to which data is being fit. There are few other assumptions or restrictions related to the data

or the specific method employed for clustering. The literature on cluster analysis is rich and various reviews are available (Jain, Murty, and Flynn 1999).

Various researchers in information systems have used cluster analysis to either explore data patterns or to confirm *a priori* data groupings. Typically, data groupings are used in further analysis to test hypotheses related to other variables. For example, in a recent study, Ravichandran and Rai (1999) studied total quality management in information systems development. Through a review of the quality and systems development literature, they developed eleven quality management constructs and two quality performance constructs. Each scale was empirically validated using confirmatory factor analysis. Next, the scales were examined through cluster analysis to examine patterns of association. Three clusters were selected based on a non-hierarchical cluster procedure – high TQM experience and success, low TQM experience service firms, and low TQM manufacturing firms. Next the fourteen variables were tested across the three clusters for differences.

In a second study, Segars and Grover (1999) examine profiles of strategic information systems planning. They identified six dimensions of strategic planning – comprehensiveness, formalization, focus, flow, participation, and consistency. Each dimension is validated using confirmatory analysis and then a cluster analysis is initiated to create profiles based on the six dimensions. Five clusters emerge and identified and labeled as the Design School, Planning School, Positioning School, Learning School, and the Political School. Each cluster is next compared using scales of planning effectiveness.

In both the above example, interpretations are that responding organizations fall into one and one cluster only. In the case of the second example, one organization may belong to either the Design School or the Planning School but could not have memberships or characteristics of both profiles. This restriction is useful for obtaining pure taxonomic maps but may not adequately reflect the practice of information systems planning.

One alternative to crisp clustering that relaxes this condition is fuzzy clustering which allows for memberships in all clusters and calculates a membership grade in each cluster. As a result, further analysis may be at the cluster level or at the individual level. Below, we provide the technical details of fuzzy clustering and demonstrate the method empirically using questionnaire data in the field of end-user computing.

The Fuzzy Clustering Algorithms

The problem of cluster analysis is well known - its goal is to find the best partition of n entities into c classes. The most well known example of this approach is the fuzzy c-means method initially proposed by Dunn (1973) and generalized by Bezdek (1981). The fuzzy c partition is defined in such a way that the membership of an entity to a cluster expresses a part of the cluster's prototype reflected in the entity. Thus, an entity may bear 60% of a prototype A and 40% of the prototype B, which simultaneously expresses the entity's membership to the respective clusters. The prototypes are considered as offered by the knowledge domain. In this case the prototype are the three end-user functions – operations, development, and control.

The Fuzzy c-Means Algorithm

The fuzzy c-means (FCM) algorithm (Bezdek 1981) is one of the most widely used methods in fuzzy clustering. It is based on the concept of fuzzy c-partition (Ruspini 1969), summarized as follows.

Let X = {x1,...,xn} be a set of given data, where each data point x_k ($k=1,...,n$) is a vector in $\Re^P$, U_{cn} be a set of real c x n matrices, and c be an integer, $2 \le c < n$. Then, the fuzzy c-partition space for X is the set:

$$M_{fcn} = \left\{ U \in U_{cn} : u_{ik} \in [0,1] \right.$$
$$\left. \sum_{i=1}^{c} u_{ik} = 1, 0 < \sum_{k=1}^{n} u_{ik} < n \right\} \quad (1)$$

where u_{ik} is the membership value of xk in cluster $i(i=1,...,c)$. The value of c is assumed to be known. The aim of the FCM algorithm is to find an optimal fuzzy c-partition and corresponding prototypes minimizing the objective function:

$$J_m(U,V;X)=\sum_{k=1}^{m}\sum_{i=1}^{c}(u_{ik})^m\|x_k-v_i\|^2. \quad (2)$$

In (2), V = (v_1,v_2,...,v_c) is a matrix of unknown centers (prototypes) $v_i \in \Re^P$, $\|\cdot\|$ is the Euclidean norm, and the weighting exponent m in [1, ∝] is a constant that influences the membership value. The FCM clustering criterion belongs to the class of least squares clustering criteria (Jain 1988). To minimize criterion J_m, under the fuzzy constraints defined in (1), the FCM algorithm is defined as an alternating minimization algorithm. First, a value for c, m, and ϵ, a small positive constant is chosen; then, generate randomly a fuzzy c-partition U_0 and set iteration number $t = 0$. In the second step, use the membership values $u_{ik}^{(t)}$ to calculate the cluster centers $v^{(t)}(i=1,...,c)$ as follows:

$$v_i^{(t)}=\frac{\sum_{k=1}^{n}\left(u_{ik}^{t}\right)^m x_k}{\sum_{k=1}^{n}\left(u_{ik}^{(t)}\right)^m}. \quad (3)$$

Given the new cluster center $v_i^{(t)}$, update membership values $u_{ik}^{(t)}$:

$$u_{ik}^{(t+1)}=\left[\sum_{j=1}^{c}\left(\frac{\|x_k-v_i^{(t)}\|^2}{\|x_k-v_j^{(t)}\|^2}\right)^{\frac{2}{m-1}}\right]^{-1}. \quad (4)$$

The process stops when $\left|U^{(t+1)}-U^{(t)}\right|\le\varepsilon$, or a predefined number of iterations is reached. U^0 may be random generated or a hard cluster solution may be used to initialize membership.

One of the problems with fuzzy clusters, similar to the one encountered in hard clustering, is that of determining the number of clusters, or otherwise stated, how well does the algorithm identify structure that is present in the data – the cluster validity problem (Bezdek 1981). Multiple cluster validity functions have been developed to measure the quality of clustering. These include partition coefficients, classification entropy, proportion, exponent. One of the most well studied validity function is the compactness and the separation validity function, defined as follows:

$$S(c)=\sum_{k=1}^{n}\sum_{j=1}^{c}(u_{ik})^m\left(\|x_k-v_i\|^2-\left\|v_i-\bar{x}\right\|^2\right)$$

The smaller the value of S(c), the better the compactness and separation between the clustering groups of in-cluster samples. Therefore, one goal of the clustering algorithm is to minimize the value of S(c).

Example Analysis: End-User Computing Classification

A key research issue in the field of end-user computing (EUC) concerns learning more about those individuals who develop and use their own software applications. Descriptive studies (Benson 1983; Rivard and Huff 1985) have profiled end-users based on their background, source of their computer education, and their experience. Classification studies (Rockart and Flannery 1983) develop typologies of end-users and categorize them into one of six different categories ranging from "non-programming users" to "IS programmers." Similarly, Cotterman and Kumar (1989) develop a framework in which users are classified based on the dimensions of EUC - operations, development, and control.

In most studies, end-users are classified into one of the several categories through self-selection (Schiffman, Meile, Igbaria 1992; Barker and Wright 1997). An alternative to the self-selection mechanism could be the operationalization of the specific dimensions of a typology and using procedures such as cluster analysis to classify end-users. However, data-analytic approaches (i.e., cluster analysis) that are applied to understanding user populations generally do not reflect the conditions where end-users may be performing a variety of tasks (e.g., data access and application development) but to differing degrees.

Fuzzy clustering enables us to better understand the classification of end-users. We adopt classification dimensions as described by Cotterman and Kumar (1989) – operations, development, and control. These three dimensions are then used to describe a procedure where end-users may be shown to have differing memberships in each of the fuzzy clusters developed. Thus, membership (or non-membership) of a subject in one (non-overlapping) or multiple (overlapping) clusters is replaced by gradual membership, indicating the probability or the degree of membership of an end-user in a cluster.

Data Collection, Measurement, and Procedures

Data was collected from 196 end-users through a questionnaire survey sent to a random sample of 250 employees in a large Midwestern city in the United States. The original sample was obtained from the alumni list of the business school of one of the researchers. The average age of the 196 respondents was 32 years, and their average tenure with their organizations was 7.3 years. There were seventy females and 125 male employees in the sample. The distribution of the industries for the respondents is as follows: 20 (10.5%) banking and finance; 44 (23.0%) consulting; 7 (3.7%) retail; 14 (7.3%) manufacturing; 10 (5.2%) insurance; 17 (8.9%) health care; 14 (7.4%) other services; 41 (21.5%) communications; 22 (11.5%) transportation; and 7 (1%) other misc.

End-User Computing Dimensions

The empirical use of Cotterman and Kumar's (1989) taxonomy is based on the measurement of three dimensions of end-user computing – operation, development, and control. These dimensions were operationalized based on Govindarajulu (2001).

Operation was measured by an average of the extent to which a respondent used different types of applications - level-1, level-2, and level-3 applications. Examples of Level-1 applications are Powerpoint presentations, static web pages and simple applications. Level-2 applications are spreadsheet applications using financial or statistical formulations, macros, database applications using SQL queries, dynamic web pages which use JAVA/Perl, VB, VBS, and programs using GUI-based Fox Pro/C++. Level-3 applications are complex programs that involve extensive use of advanced features of COBOL, or GUI languages such as VB/Fox Pro/C++, CAD/CAM, etc.

Development was measured by a four-item scale, which assesses the extent to which end-users are involved in the development tasks, for example, specifications of end user requirements, the design of applications, actual programming of applications, and systems implementations.

Control is measured by a five-item scale which assesses the authority of the end user to acquire hardware, software, initiate, manage, and implement end-user applications, collect, store, and use data for end-user applications, and assign, acquire, and staff development of end-user applications.

Results and Conclusion

The four development items loaded to a single factor with an alpha value of 0.92. The five control items also loaded to a single factor with an alpha value of 0.91. Scores on operation, development, and control were standardized and subjected to a fuzzy cluster analysis.

U_0 was estimated with results from a hard cluster analysis. A SAS IML algorithm was written to implement the alternating optimization algorithm as described above. Data was clustered into 2,3,4, and 5 clusters using a fuzzy exponent value of 1.3. The algorithm converged after 20, 21, 25, and 30 iterations respectively with a ϵ value of 0.0001.

The value of various validity functionals – Partition Coefficient, Non Fuzzy Index, and the Fuzzy Performance Index – is calculated for 2,3,4, and 5 cluster solutions. There is consistent indication with all validity functions of a three-cluster solution. Thus, the three-cluster solution is selected for qualitative interpretation.

The cluster centers for the three clusters are as follows: cluster 1- 4.2, 5.3, 3.4; cluster 2 – 3.9, 6.0., 1.1; and cluster 3 – 3.8, 5.7., 1.3. Under a crisp-cluster interpretation, cluster 2 can be interpreted as the User-Developer-Controller (UDC) cluster; cluster 3 the User (U) cluster; and cluster 1 the User-Developer (UD) group. In addition to the cluster centers, cluster membership is calculated for each observation, thus respondent 1 can best be represented as falling 4.1% in the user category, 85.3% in the developer category, and 10.4% in the controller category. Thus, finer details may be obtained using fuzzy clustering than would be possible within the hard cluster framework. However, a crisp cluster solution may be imposed on the data by classifying each observation into one cluster. For example, given the data above, respondent 1 would be classified into the cluster 2 (membership in cluster 2 changed to 1) and membership reduced to zero in other clusters. Cluster centers are then recalculated given the new membership data and would be different.

There are several implications of benefits of conducting a fuzzy cluster analysis. Typically, cluster analysis leads to further testing of hypotheses regarding differences between clusters. For example, we might be interested in the area of EUC support in the case of the above example. Various studies have theorized that different types of end-users require different types of support, i.e., higher support needs are required in cases where end-users perform more development tasks. Similarly, end-users who perform functions related the control dimension, may require higher support needs. In a crisp cluster analysis, this expectation would be examined as follows:

1. develop and validate measures of end-user support; and
2. perform comparison tests of EUC support across different clusters.

While similar analyses may be performed in the case of fuzzy clustering, the expectation of higher support needs may be examined at different levels. At the *cluster level,* comparison tests may be performed by imposing a crisp structure on fuzzy membership data. In addition, an analysis might be performed at the *individual observation level*. Thus, correlations between increasing membership within a cluster and support needs might be examined. Finally, clusters might be combined to conduct, what we term a cluster-conjunctive analysis. Here the consideration is the examination of the form of the model used to predict level of support need based on cluster memberships.

In general, several methods are available to test for the form of this relationship. Most of the methods fall within the purview of mixture models and reviewed in Cornell (1981). These models may take one of the following forms – linear, quadratic, full cubic, and special cubic. A sequential fitting strategy is followed using the lack-of-fit significance as the criteria.

Thus, we find that fuzzy clustering offer an alternative mechanism to crisp cluster analysis and may be used to develop more detailed studies in information systems.

References

Barker, R.M., and Wright, A.L., "End User Computing Level, Job Motivation and User Perceptions of Computing Outcomes: A Field Investigation*," SIGCPR 97*, San Francisco, pp. 224-233.

Bensen, D.H., "A Field Study of End User Computing : Findings and Issues," MISQ (7:4), 1983, pp. 35-45.

Bezdek., J.C., Pattern Recognition With Fuzzy Objective Function Algorithms, New York: Plenum Press, 1981.

Cornell, J.A., Experiments With Mixtures, New York: John Wiley, 1981.

Cotterman, W.W., and Kumar, K., "User Cube: A Taxonomy of End Users," Communications of the ACM," (32:11), 1989, pp. 1313-1320.

Govindarajulu, C., "End Users: Who They Are," Communications of the ACM," Forthcoming, 2001.

Jain, A.K., Algorithms for Clustering Data, Upper Saddle River, NJ: Prentice Hall, 1988.

Jain, A.K., Murty, M.N., and Flynn, P.J., "Data Clustering: A Review," ACM Computing Survey, (31: 3), September, 1999, pp. 264-323.

Ravichandran, T., and Rai, Arun, "Total Quality Management in Information Systems Development: Key Constructs and relationships," Journal of Management Information Systems, Winter 1999-2000, (16:3), pp. 119-155.

Rivard, S., and Huff, S.L., "User Developed Applications: Evaluations of Success From the DP Department Perspective," MISQ, (8:1), 1984, pp. 39-49.

Ruspini, E.H., "A New Approach to Clustering," Information Control, 1969, pp. 22-32.

Schiffman, S.J., Meile, L.C., and Igbaria, M., "An Examination of End-User Types, Information & Management, (22), 1992, pp. 207-215.

Segars, A.H., and Grover, V., "Profiles of Strategic Information Systems Planning, Information Systems Research, (10:3), September, 1999, pp. 199-232.