World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:1, 2007

Open Science Index, Computer and Information Engineering Vol:1, No:1, 2007 publications.waset.org/13836.pdf

# Fuzzy Clustering of Categorical Attributes and its Use in Analyzing Cultural Data

George E. Tsekouras, Dimitris Papageorgiou, Sotiris Kotsiantis, Christos Kalloniatis, and Panagiotis Pintelas

***Abstract*****—** We develop a three-step fuzzy logic-based algorithm for clustering categorical attributes, and we apply it to analyze cultural data. In the first step the algorithm employs an entropy-based clustering scheme, which initializes the cluster centers. In the second step we apply the fuzzy *c*-modes algorithm to obtain a fuzzy partition of the data set, and the third step introduces a novel cluster validity index, which decides the final number of clusters.

***Keywords*****—**Categorical data, cultural data, fuzzy logic clustering, fuzzy *c*-modes, cluster validity index.

## I. INTRODUCTION

CLUSTERING categorical objects is an important operation in data mining. Categorical data clustering (CDC) has been investigated by many researchers. A common approach among the various CDC procedures is to use hierarchical clustering schemes, which are based on agglomerative clustering [1] or on the use of similarity [2] and disimilarity measures [3]. Ralambondrainy [4], converted multiple categorical attributes into binary attributes by using 0 or 1 for absence or presence of a category, respectively. Then he treated these binary values as real differences and used them in the well-known *c*-means algorithm. However, a major drawback related to this approach is that the produced number of binary values becomes very large when each attribute is described by many categories. To reduce the computational complexity of a CDC algorithm, Huang [5] used a simple matching dissimilarity measure and introduced the *c*-modes algorithm, which is an extension of the classical *c*-means algorithm. However, as with most clustering algorithms, the *c*-modes is very sensitive to initialization. In his later work [6], Huang generalized the *c*-modes approach by introducing the fuzzy *c*-modes. The extistence of fuzziness in a clustering process exhibits two appealing features. Firstly, it provides a flexible representation of the data structure because each object belongs to more than one cluster with different degrees of participation. Secondly, it is able to model the uncertainty typically involved in a data set. Despite the fact that fuzzy *c*-modes is a very fast and very efficient method, it suffers from two major problem: (a) it is very sensitive to initialization, and (b) it requires an *a priori* knowledge of the number of clusters.

In this paper we propose a three-level hierarchical fuzzy clustering algorithm to cope with these two problems. The first problem is solved by employing an entropy-based clustering algorithm, which does not use any random guesses for the initial conditions. On the other hand, the second problem is solved by introducing a novel cluster validity index for CDC. Finally, the algorithm is used to detect and analyse shifting patterns, which are related to the cross-cultural adaptation of individuals in a specific cultural environment.

## II. FUZZY *C*-MODES

Let $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}$ be a set of categorical objects. Each object is described by a set of attributes $A_1, A_2,..., A_p$. The *j*-th attribute $A_j\ (1\le j\le p)$ is defined on a domain of categories denoted as $DOM(A_j)=\{a_j^1,a_j^2,...,a_j^{q_j}\}$, where $q_j$ is the number of categories assigned to $A_j$. Thus, the *k*-th categorical object $\boldsymbol{x}_k\ (1\le k\le n)$ is described as: $\boldsymbol{x}_k=[x_{k1},x_{k2},...,x_{kp}]$ with $x_{kj}\in DOM(A_j)\ (1\le j\le p)$. Let $\boldsymbol{x}_k=[x_{k1},x_{k2},...,x_{kp}]$ and $\boldsymbol{x}_l=[x_{l1},x_{l2},...,x_{lp}]$ be two categorical objects. Then, the matching dissimilarity between them is defined as [5],

$$D(\boldsymbol{x}_k,\boldsymbol{x}_l)=\sum_{j=1}^{p}\delta(x_{kj},x_{lj})\quad (1\le k\le n,\ 1\le l\le n,\ \ k\ne l) \qquad (1)$$

where

$$\delta(x_{kj},x_{lj})=\begin{cases}0, & if\quad x_{kj}=x_{lj}\\ 1, & otherwise\end{cases} \qquad (2)$$

Then, the fuzzy *c*-modes algorithm is based on minimizing the following objective function,

G. E. Tsekouras is with the Department of Cultural Technology and Communication, University of the Aegean, Faonos and Harilaou Trikoupi str., 81100, Mytilene, Greece (tel: +310-2251-0-36631, Fax: +301-2251-0-36609, e-mail: gtsek@ ct.aegean.gr).

D. Papageorgiou is with the Department of Cultural Technology and Communication, University of the Aegean, Faonos and Harilaou Trikoupi str., 81100, Mytilene, Greece (e-mail: d.papageorgiou@ ct.aegean.gr).

S. Kotsiantis is with the ESDLAB, University of Patras, Greece (tel: +301-2610-997833.

C. Kalloniatis is with the Department of Cultural Technology and Communication, University of the Aegean, Faonos and Harilaou Trikoupi str., 81100, Mytilene, Greece (e-mail: ch.kalloniatis@ ct.aegean.gr).

P. Pintelas is with the ESDLAB, University of Patras, Greece (tel: +301-2610-997313.

$$J_m(\boldsymbol{U},\boldsymbol{V}) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m D(\boldsymbol{x}_k,\boldsymbol{v}_i) \tag{3}$$

subject to the following equality constraint,

$$\sum_{i=1}^{c} u_{ik} = 1 \tag{4}$$

where $n$ is the number of categorical objects, $c$ is the number of clusters, $\boldsymbol{x}_k\ (1 \le k \le n)$ is the $k$-th categorical object, $\boldsymbol{v}_i\ (1 \le i \le c)$ is the $i$-th cluster center, $u_{ik}$ is the membership degree of the $k$-th categorical object to the $i$-th mode, $\boldsymbol{U} = \{[u_{ik}]\}$ is the partition matrix, $\boldsymbol{V} = \{[\boldsymbol{v}_i]\}$ is the set of modes (cluster centers), and $m \in (1, \infty)$ is a factor to adjust the membership degree weighting effect, usually known as fuzziness parameter.

It can be easily shown that the membership degrees that solve the above constrained optimization problem are given as,

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{D(\boldsymbol{x}_k,\boldsymbol{v}_i)}{D(\boldsymbol{x}_k,\boldsymbol{v}_j)}\right)^{1/(m-1)}} \quad (1 \le i \le c,\, 1 \le k \le n) \tag{5}$$

On the other hand, the cluster centers that optimize the objective function in (3) are derived by the following theorem,

*Theorem 1*

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ be a set of categorical objects described by categorical attributes $A_1$, $A_2$,…, $A_p$, and $DOM(A_j) = \{a_j^1, a_j^2, ..., a_j^{q_j}\}$ where $q_j$ is the number of categories that are assigned to the $j$-th attribute $A_j$ $(1 \le j \le p)$. Let the membership degrees $u_{ik}\ (1 \le i \le c,\, 1 \le k \le n)$ be fixed. Then the locations of the cluster centers (modes) that minimize the objective function in (3) are determined as: $\boldsymbol{v}_i = [v_{i1}, v_{i2}, ..., v_{ip}]$, where $v_{ij} = a_j^r \in DOM(A_j)$ with,

$$\sum_{k,\, x_{kj}=a_j^r}(u_{ik})^m \ge \sum_{k,\, x_{kj}=a_j^t}(u_{ik})^m \qquad (1 \le t \le q_j,\, r \ne t) \tag{6}$$

***Proof***

The proof of the theorem can be found in [6].

The fuzzy $c$-modes algorithm is now given as follows:

Step 1). Select a value for the parameters $m$ and ε.

Step 2). Initialize the modes $\boldsymbol{v}_i\,(1 \le i \le c)$. Using theorem 1, calculate the membership degrees $u_{ik}$, and using eq. (3) determine $J_m(\boldsymbol{U},\boldsymbol{V})$.

Step 3). Set $J_m^{old}(\boldsymbol{U},\boldsymbol{V}) = J_m(\boldsymbol{U},\boldsymbol{V})$

Step 4). Based on theorem 1, update the cluster centers $\boldsymbol{v}_i$ $(1 \le i \le c)$.

Step 5). Based on eq (5), calculate the membership degrees $u_{ik}$ $(1 \le i \le c,\, 1 \le k \le n)$.

Step 6). Update the $J_m(\boldsymbol{U},\boldsymbol{V})$.

Step 7). If $\left|J_m(\boldsymbol{U},\boldsymbol{V}) - J_m^{old}(\boldsymbol{U},\boldsymbol{V})\right| \le \varepsilon$ stop. Else go to step 3.

## III. The Proposed Method

The flow sheet of the proposed algorithm is shown in Fig. 1. More specifically, in the first step we use an entropy-based clustering scheme. This scheme is used as a data pre-processor unit in order to extract the initial number of clusters and the respective cluster centers. Thus, in this step we are not concerned with a rigorous design of the entropy-based clustering scheme. The only restriction here is that the initial number of clusters must not be either too large or too small. In the second step we use the fuzzy $c$-modes and in the third step we develop a novel validity index, which determines the optimal number of clusters. The detailed analysis of the above steps is described in the following subsections.

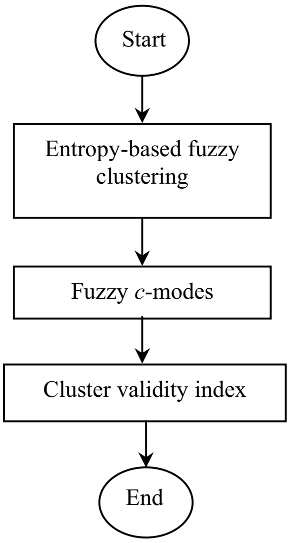

Figure 1: The flow sheet of the proposed CDC algorithm.

### A. Entropy-Based Clustering

The entropy-based clustering scheme developed in this subsection extends the algorithm proposed in [7] in order to handle categorical data, and it is described next. The entropy value between two categorical objects $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$ is given by the following relation,

$$H_{kl} = -E_{kl}\log_2(E_{kl}) - (1-E_{kl})\log_2(1-E_{kl}) \tag{7}$$

where $k \ne l$, and $E_{ij}$ is a similarity measure between $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$, and it is defined by the next formula,

$$E_{kl} = \exp\{-a\, D(\boldsymbol{x}_k,\boldsymbol{x}_l)\} \tag{8}$$

where $a$ is a design parameter, and $D(\boldsymbol{x}_k, \boldsymbol{x}_l)$ is given in eq. (1). Then, the total entropy of the vector $\boldsymbol{x}_k$ with respect to
the rest of the categorical objects is given by the next equation,

$$\boldsymbol{H}_k = -\sum_{\substack{l=1 \\ l \neq k}}^{n} \left[ E_{kl} \log_2 (E_{kl}) - (1 - E_{kl}) \log_2 (1 - E_{kl}) \right] \qquad (9)$$

Since, as mentioned previously, we are not concerned with a rigorous design of the entropy-based algorithm, a reasonable choice is to select a value for the parameter $a$ such that: $a \in (0,1)$. From eq. (9) we can point out the following useful remark: *the total entropy of a categorical object is small when many neighboring objects surround this object. Therefore, an object with a small total entropy value is a good nominee to be a cluster center.* Based on this remark the entropy-based clustering algorithm is described next:

*Entropy-Based Clustering Algorithm*

Select values for the design parameter $a \in (0,1)$ and for the parameter $\beta \in (0,1)$. Initialy, set the number of clusters equal to $c$=0.

Step 1). Using eq. (9) determine the total entropies for all data vectors $\boldsymbol{x}_k$ $(1 \leq k \leq n)$.

Step 2). Set $c$=$c$+1.

Step 3). Calculate the minimum entropy $\boldsymbol{H}_{\min} = \min_k \{\boldsymbol{H}_k\}$.
Select the object $\boldsymbol{x}_{\min}$ that corresponds to $\boldsymbol{H}_{\min}$ as the center element of the $c$-th fuzzy cluster: $\boldsymbol{v}_c = \boldsymbol{x}_{\min}$

Step 4). Remove from the set $\boldsymbol{X}$ all the categorical objects having similarity with $\boldsymbol{x}_{\min}$ greater than β and assign them to the $c$-th cluster.

Step 5). If $\boldsymbol{X}$ is empty stop. Else turn the algorithm to step 2.

The above procedure produces $c$ clusters, the centers of which are denoted as: $\boldsymbol{v}_i$ $(1 \leq i \leq c)$. This algorithm appears two major advantages: (a) it is one-pass through the data set and therefore it is a very fast procedure that is easy to implement, (b) it does not assume any initial cluster centers but rather it selects these centers based on the data structure. The clusters that are produced by the above algorithm are further elaborated by the fuzzy $c$-modes. Finally, the third step uses the cluster validity process, which is described within the next subsection.

### *B. Cluster Validity Index*

Cluster validity concerns the determination of the optimal number of clusters. In this section we develop a cluster validity index for CDC, which is based on the use of compactness and separation criteria. The variation $(\sigma_i)$, and the fuzzy cardinality $(n_i)$ of the $i$-th cluster are respectively given as [8],

$$\sigma_i = \sum_{k=1}^{n} (u_{ik})^m D(\boldsymbol{x}_k, \boldsymbol{v}_i) \quad , \quad 1 \leq i \leq c \qquad (10)$$

$$n_i = \sum_{k=1}^{n} u_{ik} \quad , \quad 1 \leq i \leq c \qquad (11)$$

Using these two concepts, the global compactness $(\pi)$ of the fuzzy partition is defined as,

$$\pi = \sum_{i=1}^{c} \frac{\sum_{k=1}^{n} (u_{ik})^m \, D(\boldsymbol{x}_k, \boldsymbol{v}_i)}{n_i} \qquad (12)$$

In order to define the fuzzy separation, the $i$-th cluster center $\boldsymbol{v}_i$ is viewed as the center of a fuzzy set, which consists of the rest of the vectors $\boldsymbol{v}_j$ $(1 \leq j \leq c,\ j \neq i)$ and its membership function is given as follows,

$$\mu_{ij} = \frac{1}{\sum_{\substack{l=1 \\ l \neq j}}^{c} \left( \frac{D(\boldsymbol{v}_j, \boldsymbol{v}_i)}{D(\boldsymbol{v}_j, \boldsymbol{v}_l)} \right)^{1/(m-1)}} \quad (i \neq j) \qquad (13)$$

Then, the fuzzy separation is defined as,

$$s = \sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} (\mu_{ij})^m \, D(\boldsymbol{v}_i, \boldsymbol{v}_j) \quad (j \neq i) \qquad (14)$$

Finally, the validity index is given as the ratio between the global compactness ($\pi$), and the fuzzy separation ($s$):

$$S = \frac{\sum_{i=1}^{c} \frac{\sum_{k=1}^{n} (u_{ik})^m D(\boldsymbol{x}_k, \boldsymbol{v}_i)}{n_i}}{\sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} (\mu_{ij})^m \, D(\boldsymbol{v}_i, \boldsymbol{v}_j)} \qquad (15)$$

The optimum number of clusters using the index $S$ is the one that corresponds to its lowest value.

## IV. Analyzing Cultural Data

One of the major fields in cross-cultural adaptation research area is the so called ethnocultural identity [9]. The impact of ethocultural identity to the cross-cultural adaptation of immigrants in a foreign cultural environment, can be measured by the following empirical indicators [10]: (a) knowledge of the host communication system, (b) cognitive complexity in responding to the host environment, (c) affective co-orientation with the host culture, and (d) behavioral capabilty to perform various interactions in the host environment. To cary out our experiment, we measured the above indicators by creating a questionnaire, which consists of four attributes (questions) namely,

$A_1$={Speak the language of the host environment with family or close friends?}
$A_2$={Listen to the music of the host environment?}
$A_3$={Read newspapers/magazines of the host environment?}
$A_4$={Friendly interaction with people of the host environment?}.

Each of the above questions has three possible answers: Never (=1), Often (=2), Always (=3). Thus, each of the above 4 attributes is assigned 3 categories.

TABLE I: OPTIMAL NUMBER OF CLUSTERS FOR VARIOUS VALUES OF THE FUZZINESS PARAMETER $m$

| | $m$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1.1 | 1.5 | 2 | 2.5 | 3 | 4 | 7 |
| $c_{\text{opt}}$ | 4 | 4 | 4 | 4 | 4 | 7 | 10 |

The relation between the above attributes and the empirical indicators mentioned previously, can be formulated as follows: the first empirical indicator is described by the attribute $A_1$, the second one by the attribute $A_3$, the third one by the attribute $A_2$, while the fourth indicator is described by the attribute $A_5$. As it is easily undestood, the above correspodence is not unique, meaning that each indicator is described by a plethora of factors and not only by the above attributes. However, we choose to use only these attributes because they are simple, easy to be undestood, and the answer is not time consuming.

TABLE II: FINAL CLUSTER CENTERS AND THE RESPECTIVE LABELS

| $i$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | Label |
|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 1 | 1 | 2 | Low (L) |
| 2 | 2 | 2 | 2 | 2 | 3 | Good (G) |
| 3 | 2 | 2 | 2 | 3 | 3 | Very Good (VG) |
| 4 | 2 | 3 | 3 | 2 | 3 | High (H) |

During the experiment, 50 immigrants who live in Greece provided answers to the above questions once per month for 12 months. The experiment took place between January 2003 and December 2003. Thus, the total number of categorical objects that were generated was equal to $N$=600.

Table 1 depicts the optimal number of clusters obtained by the algorithm for various values of the parameter $m$. Based on this table, a reasonable choice is to select the optimal number of clusters equal to $c_{\text{opt}}$=4. The final cluster centers for this selection are given in Table 2. In the next step, each fuzzy cluster is assigned a label, which corresponds to the degree of cross-cultural adaptation capability provided by the categories of the respective cluster center. These labels are shown in the most right column of Table 2. Since a specific fuzzy cluster may share common objects with a specific month, we assign to it weight values that express the significances of this cluster with respect to each month. These weights are determined by the number of objects that belong both to the this cluster and each month, and are calculated as ,

$$\pi_i^l = \sum_{x_k \in \boldsymbol{X}_l} u_{ik} \Bigg/ \sum_{i=1}^{c} \sum_{x_k \in \boldsymbol{X}_l} u_{ik} \quad (1 \le i \le c, 1 \le l \le 12) \qquad (16)$$

where $\boldsymbol{X}_l$ is the number of objects that belong to the $l$-th month, with $\boldsymbol{X} = \bigcup_{l=1}^{12} \boldsymbol{X}_l$. Fig. 2 shows the shifting patterns of the cross-cultural adaptation data. From this figure we can see that the pattern “Low” decreases as time passes, the pattern “Good” remains almost constant. On the other hand, the patterns “Very Good” and “High” increase their weights of significance, which means that as time passes the adaptation of the individuals becomes more and more efficient.

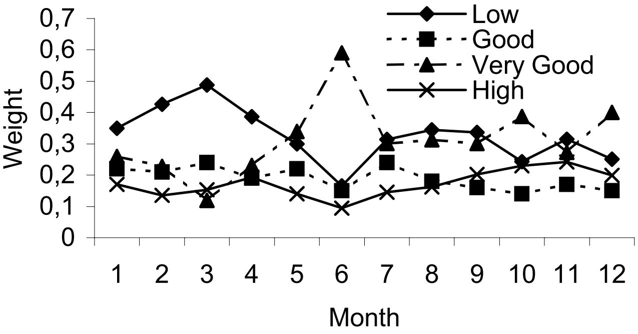

Figure 2: Shifting patterns as a function of time.

## V. CONCLUSIONS

This paper presented the development and evaluation of a CDC algorithm. The proposed algorithm is able to sufficiently reduce the dependence of the clustering process on initialization. Moreover, the algorithm is equiped with a novel cluster validity index to determine the optimal number of clusters. The algorithm was used to study cross-cultural data, where shifting patterns that correspond to certain levels of adaptation capability of individuals in a specific foreign cultural environment, were detected and analyzed.

## REFERENCES

[1] T. Morzy, Wojciechowski M., & Zakrzewicz M., Scalable hierarchical clustering method for sequences of categorical values, *Lecture Notes in Artificial Intelligence*, 2035, 2001, 282-293.

[2] S. Guha, R. Rastogi, & K. Shim, ROCK: A robust clustering algorithm foa categorical attributes, *Information Systems*, 25(5), 2000, 345-366.

[3] K. Mali, & M. Sushmita, “Clustering of symbolic data and its validation, *Lecture Notes in Artificial Intelligence*, 2275, 2002, 339-344.

[4] H. Ralambondrainy, A conceptual version of the k-means algorithm”, *Pattern Recognition Letters*, 16, 1995, 1147-1157.

[5] Z. Huang, Extensions of the k-means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery*, 2, 1998, 283-304.

[6] Z. Huang, & M. K. Ng, A fuzzy k-modes algorithm for clustering categorical data, *IEEE Transactions on Fuzzy Systems*, 7(4), 1999, 446-452.

[7] J. Yao, M. Dash, S. T. Tan, and H. Liu, Entropy-based fuzzy clustering and fuzzy modeling, *Fuzzy Sets and Systems*, 113, 2000, 381-388.

[8] L. X. Xie, G. Beni, A validity measure for fuzzy clustering, *IEEE Trans. PAMI*, 13, 1991, 841-847.

[9] K. Cushner, R.W. Brislin, *Improving intercultural interactions: Modules for training programs*, Vol. 2, Sage Publications, 1997.

[10] Y.Y. Kim, Communication and cross-cultural adaptation: an integrative theory, Multilingual Matters Ltd, England, 1988.