A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional

Sotirios P. Chatzis
Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, South Kensington Campus SW7 2BT, UK

Keywords: Categorical data; Fuzzy clustering; Fuzzy c-means; Regularization; Gaussian-Multinomial assumption

ABSTRACT

Gath–Geva (GG) algorithm is one of the most popular methodologies for fuzzy c-means (FCM)-type clustering of data comprising numeric attributes; it is based on the assumption of data deriving from clusters of Gaussian form, a much more flexible construction compared to the spherical clusters assumption of the original FCM. In this paper, we introduce an extension of the GG algorithm to allow for the effective handling of data with mixed numeric and categorical attributes. Traditionally, fuzzy clustering of such data is conducted by means of the fuzzy k-prototypes algorithm, which merely consists in the execution of the original FCM algorithm using a different dissimilarity functional, suitable for attributes with mixed numeric and categorical attributes. On the contrary, in this work we provide a novel FCM-type algorithm employing a fully probabilistic dissimilarity functional for handling data with mixed-type attributes. Our approach utilizes a fuzzy objective function regularized by Kullback–Leibler (KL) divergence information, and is formulated on the basis of a set of probabilistic assumptions regarding the form of the derived clusters. We evaluate the efficacy of the proposed approach using benchmark data, and we compare it with competing fuzzy and non-fuzzy clustering algorithms.

1. Introduction

Cluster analysis is a methodology developed for capturing local substructures in multivariate data, by applying an affinity criterion to group data points. A significant aspect of cluster analysis that comes to the fore, with increasing complexity, is the rigidity (crispness) of the partition that is looked for. Relaxing this rigidity has constituted, in the past, a domain of research in the framework of cluster analysis. Many authors have proposed a fuzzy setting as the appropriate approach to cope with this problem. Fuzzy set theory, proposed in Zadeh (1965), gives an idea of uncertainty of belonging, which is described by a membership function. Under this regard, a fuzzy clustering algorithm obtains a fuzzy c-partition u of a given dataset \( X = \{x_i\}_{i=1}^n \), where

\[
u = \{u_{ij}\}_{i,j=1}^c\n\]

The \( u_{ij} \) is called the degree of membership of the jth data point, \( x_j \), to the ith cluster, and we have

\[
0 \leq u_{ij} \leq 1 , \quad \sum_{j=1}^c u_{ij} = 1 , \quad 0 < \sum_{j=1}^c u_{ij} < n
\]

Some early applications of fuzzy set theory in cluster analysis were proposed in the works of Bellman, Kalaba, and Zadeh (1966), Ruspini (1969). During the last decades, fuzzy clustering methodologies have been widely applied as a powerful means for cluster analysis (Bezdek, Pal, Keller, & Krisnapuram, 1999). One of the most widely used approaches to fuzzy prototype-based clustering is fuzzy c-means (FCM) clustering. The FCM algorithm is an extension of the k-means algorithm (the hard c-means algorithm) and was first introduced in Dunn (1974). Bezdek (1981) provided a generalization of the FCM algorithm introducing the degree of fuzziness, taking values not less than one, and establishing a general estimation procedure with proved convergence. The FCM algorithm comprises minimization of the fuzzy objective function

\[
J_u \triangleq \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m d_{ij}
\]

to obtain the fuzzy membership functions, \( u_{ij} \), and the point-prototypes (centroids) \( \mu_i \) of the derived clusters, in a coordinate descent fashion. In the expression of the fuzzy objective function \( J_{\mu, \phi} \geq 1 \) is an weighting exponent on each fuzzy membership function \( u_{ij} \), controlling their degree of fuzziness, and is called the fuzzifier of the clustering algorithm, and \( d_{ij} \) is the dissimilarity functional of the algorithm.

In the original formulation of the FCM algorithm the dissimilarity functional \( d_{ij} \) is taken as the Euclidean distance of the jth observed data point, \( x_j \), from the centroid of the ith cluster, \( \mu_i \), i.e.,

\[
d_{ij} = |x_j - \mu_i|_2
\]

This selection implies, in essence, the assumption of data organized into spherical clusters. However, this might not be the case in many
real-life applications. To allow for the mitigation of these issues, Gustafson and Kessel (1979) proposed a modification of the FCM algorithm by employing an adaptive distance norm, to detect clusters of different geometrical shapes in one data set. Gath and Geva (1989) defined an exponential distance for the FCM algorithm to obtain a fuzzy approach to maximum-likelihood estimation of Gaussian mixture models. Based on the assumption that the clusters to be derived are of the form of a Gaussian likelihood function, they introduce a suitable dissimilarity function $d_{ij}$ the FCM algorithm that effectively incorporates this prior assumption regarding the clustered data likelihoods into the fuzzy clustering procedure. This is effected by setting the dissimilarity functional $d_{ij}$ of the FCM algorithm as the negative log-likelihood of the Gaussian distribution, i.e.,

$$d_{ij} = -\log p(X_i | \mu_j, \Sigma_j)$$  \hspace{1cm} (5)

where $p(X_i | \mu_j, \Sigma_j)$ is a Gaussian with mean $\mu_j$ (centroid of the corresponding cluster), and covariance matrix $\Sigma_j$. Recently, an extension of the Gath–Geva algorithm to incorporate the multivariate Student’s-$t$ distribution was proposed by Chatzis and Varvarigou (2008), yielding the FSM algorithm. FSM, exploiting the harder tails of the multivariate Student’s-$t$ distribution, has been shown to be more effective than the original Gath–Geva algorithm in clustering data sets containing outliers (which is usually the case with real-world data sets), yielding enhanced clustering performance for comparable computational costs.

The introduction of the fuzzifier $\gamma$ in the context of the hard $c$-means algorithm, yielding the FCM objective function (3), has been considered as a rather unnatural device by several researchers. To address this issue, Li and Mukaidono (1995) proposed a variation of the FCM algorithm introducing a new approach to fuzzy clustering by means of a maximum entropy inference (MEI) method. Miyamoto and Mukaidono (1997) reformulated the MEI approach by considering the entropy term as a regularization term. In this case, the burden of representing fuzziness is shifted to the regularization term, in the form of a weighting factor multiplying the contribution of the regularization function to the clustering criterion. More recently, a different approach to FCM-type fuzzy clustering was proposed in Honda and Ichihashi (2005), where fuzzification is attained by means of a regularization technique where a Kullback–Leibler (KL) divergence term is introduced into the fuzzy objective function. Under this setting, the regularized by KL information fuzzy c-means (KL-FCM) algorithm is obtained, with objective function

$$J_{KL} = \frac{1}{c} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} d_{ij} + \lambda \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} \log \left( \frac{u_{ij}}{p(i)} \right)$$  \hspace{1cm} (6)

where $p(i)$ is the weight of the $i$th cluster. Here, the KL divergence term works as the fuzzifier and the parameter $\lambda$ is the model’s degree of fuzziness of the fuzzy membership values. Utilizing this variant of the FCM objective function, Chatzis and Varvarigou (2009) introduced a novel FCM-type algorithm for robust clustering of high-dimensional data, shown to outperform with grace a number of competing algorithms, such as fuzzy c-varieties (FCV), and various related statistical methods based on the expectation-maximization (EM) algorithm. Despite these advances, all these algorithms are explicitly designed to handle data comprising numeric attributes. However, in many cases, observations might contain both numeric and categorical attributes, or even only categorical attributes Mclachlan and Peel (2000). This is, for example, the case in applications dealing with medical risk assessment for individuals with various health conditions. The problem of FCM-type clustering of data comprising mixed numeric and categorical attributes is rather understudied. Indeed, to our knowledge, the only solution to this problem proposed so far is the fuzzy $k$-prototypes (FKP) algorithm Bezdek et al. (1999). Let us consider a data set $x = \{x_1, \ldots, x_n\}$, with each $d$-dimensional data point $x_i$ comprising $d_1$ real-valued (numeric) attributes and $d_2 = d - d_1$ integers (categorical attributes), i.e.,

$$x_i = (x_{ik}^{d_1})_{k=1}^{d_1} : x_{i num} = (x_{i num}^{d_1})_{k=1}^{d_1} \in \mathbb{R}^{d_1}$$

$$\land x_{i cat} = (x_{i cat}^{d_2})_{k=1}^{d_2} \in \mathbb{Z}^{d_2}$$

The FKP algorithm allows to cluster such data by using the same objective function (3) as FCM, but employing the dissimilarity functional

$$d_{ij} = \left\| x_{i num} - \mu_{j num} \right\|^2 + \sum_{k=1}^{d_2} (\delta(x_{i kat}, \mu_{j kat}))$$  \hspace{1cm} (7)

where $\mu_j$ is the centroid of the $j$th derived cluster, and

$$\delta(x, \beta) = \begin{cases} 1, & \text{if } x \neq \beta \\ 0, & \text{if } x = \beta \end{cases}$$  \hspace{1cm} (8)

In this paper, we propose a novel approach toward FCM-type fuzzy clustering of data comprising mixed numeric and categorical attributes, by making a set of suitable probabilistic assumptions regarding the form of the derived clusters, and introducing the effective means to incorporate these probabilistic assumptions into the fuzzy clustering algorithm. In this way, an extension of the Gath–Geva clustering algorithm is essentially obtained, allowing for a more flexible fuzzy modeling of data with mixed-type attributes. Our novel approach is formulated on the basis of the KL-FCM variant of the FCM algorithm, in order to exploit the intuitive advantages it offers over the conventional FCM optimization scheme, as explained previously.

The remainder of this paper is organized as follows: In Section 2, the proposed approach towards FCM-type clustering of data with mixed numeric and categorical attributes is formulated, and the updating equations of the resulting fuzzy clustering algorithm are derived. In Section 3, the effectiveness of the proposed approach is experimentally assessed using a set of benchmark data sets, and compared with the performance of competing fuzzy and non-fuzzy clustering algorithms. Finally, the concluding section summarizes our results.

2. Proposed approach: the KL-FCM-GM algorithm

2.1. Model formulation

Let us consider a data set $x = \{x_1, \ldots, x_n\}$, with each $d$-dimensional data point $x_i$ comprising $d_1$ real-valued attributes and $d_2 = d - d_1$ integers, i.e.,

$$x_i = (x_{ik}^{d_1})_{k=1}^{d_1} : x_{i num} = (x_{i num}^{d_1})_{k=1}^{d_1} \in \mathbb{R}^{d_1}$$

$$\land x_{i cat} = (x_{i cat}^{d_2})_{k=1}^{d_2} \in \mathbb{Z}^{d_2}$$

We begin our analysis with the assumption that this data set derives from a set of $c$ clusters; for the sought clusters, we proceed by making an appropriate probabilistic assumption: we consider that each cluster has the form of a suitable probability distribution. Following the related literature (e.g., Hunt & Jorgensen (2003)), a proper selection for the distribution (form) of the sought clusters is given by

$$p(x_i | \Theta) = \gamma (x_{i num} | \mu_j, \Sigma_j) \cdot \text{Mult}(x_{i cat} | \gamma_j)$$  \hspace{1cm} (9)

As we observe from (9), under the postulated probabilistic formulation, the probability density function of an observed data point $x_i$ deriving from the $i$th cluster is the product of two densities: a Gaussian $\gamma (x_{i num} | \mu_j, \Sigma_j)$ with mean $\mu_j$ and covariance $\Sigma_j$, modeling
the numeric part of the observation, and a Multinomial distribution \( \text{Mult}(x, \text{cat}|y_i) \) modeling the categorical part of the observation, with

\[
\text{Mult}(x, \text{cat}|y_i) = \prod_{k=d_{i1}+1}^{d} \prod_{l=1}^{L_k} \gamma_{ikl}^{1-\beta(x_{ikl})}
\]

where \( \gamma_{ikl} \) is the probability of the kth (categorical) attribute of one observation being equal to the lth possible value of this attribute, given the observation derives from the ith cluster, \( L_k \) is the number of possible values of the kth attribute, \( y_i = (\gamma_{ikl})_{k,l} \), \( \delta(\cdot, \cdot) \) is defined in (8), and \( \Theta_i = (\mu_i, \Sigma_i, \gamma_i) \).

Under these assumptions, we seek a fuzzy partition of the modeled data set into the postulated clusters. For this purpose, we prefer to develop a fuzzy clustering algorithm based on the KL-FCM objective function (6), as it appears to offer a more intuitive formulation of the fuzzy clustering algorithm, as discussed previously. Having selected the form of the optimized fuzzy objective function, derivation of the sought fuzzy clustering algorithm reduces to a proper selection of the dissimilarity functional of the algorithm; we need to introduce a suitable dissimilarity functional for our FCM-type algorithm that manages to effectively incorporate the information regarding the a priori assumptions about the probabilistic form of the sought clusters into the fuzzy clustering procedure.

To achieve this, we follow an approach similar to the approach adopted by the Gath–Geva algorithm Gath and Geva (1989), as well as a number of recent works on FCM-type clustering (Chatzis & Varvarigou, 2008, Chatzis & Varvarigou, 2009, Honda & Ichihashi, 2004, Honda & Ichihashi, 2005). Indeed, following Hathaway (1986), a suitable function meeting these requirements and, hence, providing a proper evaluation of the dissimilitude between a data point, say \( x_i \), and a considered, say the ith postulated cluster, taking under consideration the hypothesized probabilistic properties of the sought clusters, is the negative log-likelihood of the ith postulated cluster with respect to the data point \( x_i \), i.e., the negative log of the probability density function \( p(x_i|\Theta_i) \). This selection of the dissimilarity function has repeatedly been considered by various researchers, as the effective means to incorporate explicit information about the considered data objects’ distributions into the fuzzy clustering procedure (see e.g., Chatzis & Varvarigou, 2008, Chatzis & Varvarigou, 2009, Honda & Ichihashi, 2004, Honda & Ichihashi, 2005). Hence, we select the dissimilarity functional of our algorithm to be given by

\[
d_{ij} = - \log p(x_i|\Theta_j).
\]

Then, from (6) and (11) we obtain that the fuzzy objective functions optimized by the proposed algorithm comprise an infinite family given by

\[
J_i = - \sum_{j=1}^{c} \sum_{u=1}^{n} \log p(x_i|\Theta_j) + \sum_{j=1}^{c} \sum_{u=1}^{n} u_j \log \left( \frac{u_j}{n} \right)
\]

where \( u_j \) are the fuzzy membership functions, \( n \) is the weight of the ith cluster and, \( i \) is the degree of fuzziness of the algorithm. The FCM-type fuzzy clustering algorithm for data with mixed numeric and categorical attributes with objective function (12), based on the probabilistic assumptions regarding the form of the derived clusters given by Eqs. (9)–(11), shall be referred to as the KL-FCM algorithm for Gauss-Multinomial-distributed data (KL-FCM-GM).

2.2. Derivation of KL-FCM-GM algorithm updates

The KL-FCM-GM algorithm, as an FCM-type algorithm, is an iterative procedure, where in each iteration the employed objective function (12) is optimized over the fuzzy membership functions \( u_{ij} \), the cluster parameters \( \Theta_i \), and the cluster weights \( \pi_i \). Let us consider the \((t + 1)\)th iteration of the algorithm. To derive the updating equations of the KL-FCM-GM algorithm, we begin by considering the expression of the algorithm dissimilarity functional. We have

\[
d_{ij}(\Theta_i) = \frac{1}{2} \log |\Sigma_j| + \frac{d_1}{2} \log (2\pi) + \frac{1}{2} \text{tr} (\Sigma_j^{-1} (x_{ij} - \mu_j)^T (x_{ij} - \mu_j))
\]

Subsequently, we proceed to the estimation of the fuzzy membership functions \( u_{ij} \). Minimizing \( J_i \), over \( u_j \) under the constraint \( \sum_{j=1}^{c} u_{ij} = 1, \forall j = 1, \ldots, n \), yields (see Appendix A)

\[
u_{ij}^{t+1} = \frac{\pi_i^{t} e^{-\frac{1}{2} d_{ij}^2}}{\sum_{j=1}^{c} \pi_i^{t} e^{-\frac{1}{2} d_{ij}^2}}
\]

Concerning the cluster weights, minimizing \( J_i \), over \( \pi_i \) under the constraint \( \sum_{j=1}^{c} \pi_i = 1 \), we obtain (see Appendix A)

\[
\pi_i^{t+1} = \frac{\sum_{j=1}^{c} u_{ij}^{t+1}}{n}
\]

Minimizing \( J_i \) w.r.t. the means of the Gaussian part of the cluster likelihoods, \( \mu_j \), we obtain (see Appendix A) the estimates

\[
\mu_j^{t+1} = \frac{\sum_{i=1}^{n} x_{ij} u_{ij}^{t+1}}{\sum_{j=1}^{c} u_{ij}^{t+1}}
\]

Similar, the update of the covariances of the Gaussian part of the cluster likelihoods can be obtained by minimizing \( J_i \) w.r.t. \( \Sigma_j \). It can be proved (see Appendix A) that this yields

\[
\Sigma_j^{t+1} = \frac{\sum_{i=1}^{n} x_{ij} u_{ij}^{t+1} \Sigma_{ij}^{t+1} x_{ij} u_{ij}^{t+1} - \mu_j^{t+1} \mu_j^{t+1}}{\sum_{j=1}^{c} \sum_{i=1}^{n} x_{ij} u_{ij}^{t+1}}
\]

Finally, minimizing \( J_i \) w.r.t. \( \gamma_{ij} \), we obtain the expression for their updates, which can be proved to be given by (see Appendix A)

\[
\gamma_{ij}^{t+1} = \frac{\sum_{i=1}^{n} x_{ij} u_{ij}^{t+1} \delta(x_i, l)}{\sum_{j=1}^{c} \sum_{i=1}^{n} x_{ij} u_{ij}^{t+1}}
\]

Using these results, the KL-FCM-GM algorithm for clustering of data with mixed numeric and categorical attributes consists in the iterative procedure summarized in Algorithm 1.

Algorithm 1. KL-FCM-GM algorithm for clustering of data with mixed numeric and categorical attributes.

Set the objective function convergence threshold of the algorithm to \( W \), and the maximum number of iterations to \( T \). On the \((t + 1)\)th iteration do:
1. Compute the updates of the set of fuzzy membership functions of the algorithm, \( u = \{u_{ij}^{t+1}\} \), using (14).
2. Update the mixing weights of the postulated clusters, \( \pi_i^{t+1} \), using (16).
3. Derive the updated estimates of the cluster distribution parameters, \( \Theta_i^{t+1} \), using (17)–(19).
4. In case of algorithm convergence, i.e., \( J_i(\Theta^{t+1}) \), \( u^{t+1} - J_i(\Theta^{t+1}) \), \( u^{t+1} < W \), or if we have reached the maximum number of iterations, \( t = T \), exit; otherwise \( t = t + 1 \) and return to Step 1.
2.3. Optimal model selection techniques

From the preceding presentation, we observe that model selection for the KL-FCM-GM algorithm consists in selection of (i) a proper value for the degree of fuzziness of the clustering algorithm; and (ii) the optimal number of clusters, $c$, in the considered data sets.

Choosing the value of the degree of fuzziness parameter in FCM-type fuzzy clustering is usually conducted heuristically using a performance-based criterion, since a sound theoretical framework for its data-driven determination, by means of optimization of some kind of an objective function, has yet to be proposed (Bezdek et al., 1999). Typically, it is suggested to use values in the interval $[1,3]$, with values around 1.5 typically bearing the best results in terms of obtained error rates and generalization capacity of the trained models (see, e.g., the relevant discussions in Bezdek et al. (1999)). At this point, we would like to underline that fuzzy clustering algorithms are not sensitive to small fluctuations in the value of their degree of fuzziness, and, hence, no exhaustive search in the aforementioned interval is required for determining a good degree of fuzziness value by application of a performance-based criterion. Nevertheless, providing a sound theoretical framework for “optimal” degree of fuzziness selection for FCM-type algorithms (including the proposed algorithm as a special case), using some kind of an objective criterion, is clearly beyond the scope of this paper. As a final comment, we would like to note that cross-validation techniques, based on evaluation and comparison of the fuzzy objective function (12) of the proposed algorithm, cannot be applied for degree of fuzziness selection, since comparing the values of the fuzzy objective function obtained for different values of the degree of fuzziness is rather meaningless.

Regarding selection of the appropriate number of clusters, the alternative methods towards the solution of this problem in the context of fuzzy clustering algorithms can be classified into two categories: cross-validation techniques (Weiss & Kulikowski, 1991), and performance-based determination (Backer & Jain, 1981). Cross-validation techniques include, basically, $k$-fold cross-validation, leave-one-out, and leave-$v$-out techniques. In $k$-fold cross-validation, a given training data set is divided into $k$ subsets of (approximately) equal size. Further, a clustering algorithm is executed $k$ times, each time leaving out one of the subsets from training, but using only the omitted subset to compute the objective function of the algorithm as the model selection criterion. If $k$ equals the sample size, this is called leave-one-out cross-validation. Leave-$v$-out is a more elaborate and expensive version of cross-validation that involves leaving out all possible subsets of $v$ cases. Alternatively, performance-based determination can also be applied. Performance-based model selection comprises assignment of some task to a 10 clustering algorithm for various numbers of clusters, $c$, and selection of that value of $c$ that optimizes the algorithm performance.

3. Experimental evaluation

Here, we provide a thorough experimental evaluation of the clustering performance of our algorithm using some benchmark datasets. The considered datasets are obtained from the UCI repository (Asuncion & Newman, 2007). To assess the quality of the obtained clustering results, we utilize the provided groundtruth labeling of the used data, and we employ a number of objective criteria that measure the overlap between an obtained clustering and the corresponding groundtruth classifications. Specifically, let us consider a dataset comprising $c$ classes and its clustering obtained by application of a considered clustering algorithm. Let $\eta_i$ denote the number of data points correctly assigned to the $i$th class by the applied clustering algorithm, let $\zeta_i$ denote the data points that are incorrectly assigned to the $i$th class, and $\xi_i$ denote the data points which were incorrectly rejected from the $i$th class. Then, the precision $p_i$ of the considered clustering algorithm on the $i$th class of the clustered dataset is defined as

$$p_i = \eta_i / (\eta_i + \xi_i)$$

while the recall $r_i$ for the $i$th class is defined as

$$r_i = \eta_i / (\eta_i + \zeta_i)$$

Based on these two measures, the microprecision (micro-$p$) and the microrecall (micro-$r$) of an evaluated clustering algorithm can be derived, that is the average of the $p$ and $r$ indexes over the $c$ classes; it can be shown that these measures are given by Yang (1999)

$$\text{micro-}p = \frac{1}{n} \sum_{i=1}^{c} \eta_i$$

where $n$ is the size of the clustered dataset. In our experiments, micro-$p$ will be used as the objective criterion for the assessment of the performance of the evaluated algorithms.

3.1. Vote dataset

We begin our experiments considering the Vote dataset from the UCI repository. This dataset comprises purely categorical attributes; it consists of 435 data points, described by 16 categorical attributes. Data points are divided into two classes, namely Democrats (class #1), which contains 267 data points, and Republicans (class #2), which contains 168 data points. We run the KL-FCM-GM algorithm 100 times, each time with different (random) initializations, and we compute its (average) performance in terms of the number of correctly classified data points, and the resulting micro-$p$ values. Further, to obtain some comparative results, we repeat the same experimental procedure for the algorithms: (i) ROCK (Guha, Rastogi, & Kyuseok, 1999); (ii) Huang’s $k$-modes variant (Huang, 1998); (iii) fuzzy $k$-modes; (iv) the genetic fuzzy clustering algorithm of Gan, Wu, & Yang (2009); and (v) the centroid-based hierarchical algorithm of Jain & Dubes (1988).

The obtained results are provided in Table 1. A first comment we can make is that the cluster sizes obtained by the ROCK algorithm are lower than expected (and than the rest evaluated algorithms). This is due to the outlier-removal scheme that ROCK employs. Based on this observation, it is clear that the proposed approach completely outperforms the competition, offering a more than significant improvement in the obtained cluster quality. Note also that this improvement is attained without rejecting any kind of data (e.g., outliers), contrary to the ROCK algorithm.

3.2. Heart disease data

Further, we consider the Heart disease (Cleveland) dataset from the UCI repository. This collection consists of 303 samples comprising mixed categorical and numeric attributes; there are eight categorical and five numeric features. The available data points

<table>
<thead>
<tr>
<th>Method</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$\zeta_1$</th>
<th>$\zeta_2$</th>
<th>Micro-$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL-FCM-GM</td>
<td>240</td>
<td>163</td>
<td>5</td>
<td>27</td>
<td>0.93</td>
</tr>
<tr>
<td>ROCK</td>
<td>201</td>
<td>144</td>
<td>5</td>
<td>22</td>
<td>0.93</td>
</tr>
<tr>
<td>Huang’s $k$-modes</td>
<td>212</td>
<td>152</td>
<td>16</td>
<td>55</td>
<td>0.84</td>
</tr>
<tr>
<td>Fuzzy $k$-modes</td>
<td>219</td>
<td>157</td>
<td>11</td>
<td>48</td>
<td>0.87</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>215</td>
<td>157</td>
<td>11</td>
<td>52</td>
<td>0.86</td>
</tr>
<tr>
<td>Gan et al. (2009)</td>
<td>223</td>
<td>153</td>
<td>15</td>
<td>44</td>
<td>0.87</td>
</tr>
</tbody>
</table>
are classified into two classes: normal (class #1), including 164 data points, and heart patient (class #2), including 139 data points. As in the previous experiment, we run the KL-FCM-GM algorithm 100 times, each time with different (random) initializations, and we evaluate its (average) performance in terms of the number of correctly classified data points, and the resulting micro-p values. Further, to obtain some comparative results, we repeat the same experimental procedure for the algorithms: (i) ECOWEB (Reich & Fenves, 1991); (ii) Huang’s k-modes variant (Huang, 1998); (iii) fuzzy k-prototypes; and (iv) the similarity-based agglomerative clustering (SBAC) algorithm Li & Biswas (2002).

The obtained results are provided in Table 2. We observe that the proposed approach outperforms the competition yielding a slightly better performance (micro-p = 0.85).

### 3.3. Australian credit approval dataset

Finally, we consider the Australian credit approval dataset from the UCI repository. This dataset contains 690 data points comprising eight categorical and six numeric attributes. Data points are divided into two classes, approval (class #1), including 307 samples, and denial (class #2), including 383 samples. As in the previous experiments, we run the KL-FCM-GM algorithm 100 times, each time with different (random) initializations, and we evaluate its performance in terms of the number of correctly classified data points, and the resulting micro-p values. Further, to obtain some comparative results, we repeat the same experimental procedure for the algorithms: (i) Huang’s k-modes variant; (ii) fuzzy k-prototypes; and (iii) the method of Modha & Spangler (2003). In our investigations, both Huang’s k-modes, and fuzzy k-prototypes failed to produce any meaningful clustering result. The method of Modha & Spangler (2003) obtained a micro-p index equal to 0.83. Our algorithm yielded a slightly better performance (micro-p = 0.85).

### 4. Conclusions

In this paper, we proposed a novel approach for fuzzy clustering of data with mixed numeric and categorical attributes. Our novel approach is based on the postulation of an appropriate prior assumption regarding the distribution of the sought clusters, and the subsequent derivation of a suitable methodology for the effective introduction of these assumptions into the fuzzy clustering procedure. Our algorithm is formulated using a regularized by KL information fuzzy objective function, providing significant intuitive advantages over the original FCM algorithm objective function. The experimental evaluation of the pattern recognition performance of our model in a number of applications has shown that it outperforms competing fuzzy and non-fuzzy clustering algorithms for data with mixed numeric and categorical attributes.

### Appendix A. Parameters estimation

First, let us consider the derivation of the fuzzy membership function expressions. This can be attained by minimizing the objective function $J_i$ over $u_{ij}$ under the constraint $\sum_{j=1}^{p} u_{ij} = 1 \quad \forall j = 1, \ldots, n$. Introducing a Lagrange multiplier $\eta_j$ for each data point to enforce the constraint we have

$$\frac{\partial}{\partial u_{ij}} \left[ J_i - \sum_{j=1}^{p} \eta_j \left( \sum_{k=1}^{c} u_{ik} - 1 \right) \right] = 0$$

which yields [14]. The expressions of the mixing proportions can be derived by taking the minimization of $J_i$ w.r.t. $\pi_i$ under the constraint $\sum_{k=1}^{c} \pi_i = 1$. Using a Lagrange multiplier $\kappa$ to enforce the constraint we have

$$\frac{\partial}{\partial \pi_i} \left[ J_i - \kappa \left( \sum_{k=1}^{c} \pi_k - 1 \right) \right] = 0$$

which yields [16]. To derive the expression of $\mu_i$ we have to perform the minimization of $J_i$ w.r.t. $\mu_i$. From (12) and (13), we have that, the expression of $J_i$ ignoring terms not containing $\mu_i$ is equal to

$$\sum_{i=1}^{n} \sum_{j=1}^{p} u_{ij} \left[ -2\mathbf{x}_{i \text{num}}^{T} \mathbf{\Sigma}_i^{-1} \mu_i + \mu_i^{T} \mathbf{\Sigma}_i^{-1} \mu_i \right]$$

Since

$$\frac{\partial \mu_i^{T} \mathbf{\Sigma}_i^{-1} \mu_i}{\partial \mu_i} = 2\mathbf{\Sigma}_i^{-1} \mu_i$$

it is easy to derive (17). Similar, to compute the estimator of $\mathbf{\Sigma}_i$ we have to study the minimization of $J_i$ over it. From (12) and (13), we deduce that it is easier to derive the partial derivative of $J_i$ with respect to $\mathbf{\Sigma}_i^{-1}$. Then, we have

$$\frac{\partial \log |\mathbf{\Sigma}_i|}{\partial \mathbf{\Sigma}_i^{-1}} = -\mathbf{\Sigma}_i^{-1}$$

$$\frac{\partial (\mathbf{x}_{i \text{num}} - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x}_{i \text{num}} - \mu_i)}{\partial \mathbf{\Sigma}_i^{-1}} = \frac{\partial \text{trace} \left[ \mathbf{\Sigma}_i^{-1} (\mathbf{x}_{i \text{num}} - \mu_i)(\mathbf{x}_{i \text{num}} - \mu_i)^T \right]}{\partial \mathbf{\Sigma}_i^{-1}} = (\mathbf{x}_{i \text{num}} - \mu_i)(\mathbf{x}_{i \text{num}} - \mu_i)^T$$

from which, it is straightforward to obtain (18). Finally, regarding the parameters of the Multinomial distributions of the categorical part of the observable data, from (12) and (13), and under the constraint

$$\sum_{i=1}^{k} \gamma_{ik} = 1 \quad \forall k, i$$

implied by the Multinomial distribution, we have

$$\frac{\partial}{\partial \gamma_{ik}} \left[ J_i - \rho \left( \sum_{k=1}^{c} \gamma_{ik} - 1 \right) \right] = 0$$

where $\rho$ is a Lagrange multiplier, which is easy to obtain that, after doing the related algebra, yields (19).

### References


