Fuzzy clustering algorithms for mixed feature variables

Miin-Shen Yang*, Pei-Yuan Hwang, De-Hua Chen

Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li, Taiwan 32023, ROC

Received 8 December 2001; received in revised form 13 February 2003; accepted 17 February 2003

Abstract

This paper presents fuzzy clustering algorithms for mixed features of symbolic and fuzzy data. El-Sonbaty and Ismail proposed fuzzy \( c \)-means (FCM) clustering for symbolic data and Hathaway et al. proposed FCM for fuzzy data. In this paper we give a modified dissimilarity measure for symbolic and fuzzy data and then give FCM clustering algorithms for these mixed data types. Numerical examples and comparisons are also given. Numerical examples illustrate that the modified dissimilarity gives better results. Finally, the proposed clustering algorithm is applied to real data with mixed feature variables of symbolic and fuzzy data.

Keywords: Fuzzy clustering; Fuzzy \( c \)-means; Symbolic data; Fuzzy data; Mixed feature variables; Dissimilarity measure

1. Introduction

Clustering methods have been widely applied in various areas such as taxonomy, geology, business, engineering systems, medicine and image processing etc. (see [1,2,11]). The objective of clustering is to find the data structure and also partition the data set into groups with similar individuals. These clustering methods may be heuristic, hierarchical and objective-function-based etc.

The conventional (hard) clustering methods restrict each point of the data set to exactly one cluster. Since Zadeh [13] proposed fuzzy sets that produced the idea of partial membership of belonging described by a membership function, fuzzy clustering has been widely studied and applied in a variety of substantive areas. In the literature on fuzzy clustering, the fuzzy \( c \)-mean (FCM) clustering algorithms are the best-known methods (see [2,10]).

Let \( W \) be the data set \( \{X_1, \ldots, X_n\} \) in a \( d \)-dimensional Euclidean space \( \mathbb{R}^d \) with its ordinary Euclidean norm \( \| \cdot \| \). Let \( \{\mu_1, \ldots, \mu_c\} \) be \( c \) fuzzy sets on \( W \) with \( \sum_{i=1}^{c} \mu_i(X) = 1 \) for all \( X \) in \( W \). In this case, the fuzzy sets \( \mu_i, i = 1, \ldots, c \) construct a fuzzy \( c \)-partition of \( W \). \( \mu_i(X) \) presents the membership of the data point \( X \) which belongs to cluster \( i \). The FCM clustering is an iterative
algorithm using the necessary conditions for a minimizer \((\mu^*, A^*)\) of the objective function \(J_{FCM}\) with

\[
J_{FCM}(\mu, A) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^m \|X_j - A_i\|^2, \quad m > 1,
\]

where \(\{\mu_1, \ldots, \mu_c\}\) with \(\mu_{ij} = \mu_i(X_j)\) is a fuzzy \(c\)-partition and \(\{A_1, \ldots, A_c\}\) is the set of \(c\) cluster centers. In general, the FCM is used most for the data set in \(\mathbb{R}^d\). However, except these continuous numeric data in \(\mathbb{R}^d\), there are many other types of data such as symbolic and fuzzy types. In this paper we shall consider this FCM with the mixed types of symbolic and fuzzy data.

Symbolic data are quite different from numeric data. Symbolic variables may present human knowledge, nominal, categorical and synthetic data etc. Since 1980s, cluster analysis for symbolic data had been widely studied (see Michalski and Stepp [9], Diday [3]). In 1991–1995, Gowda et al. [5–7] proposed a hierarchical, agglomerative clustering algorithm by defining new similarity and dissimilarity measures based on “position”, “span” and “content” of symbolic objects. These defined measures presented very good results. However, the algorithm was hierarchical. On the basis of the same dissimilarity measure, El-Sonbaty and Ismail [4] created a FCM objective function for symbolic data and then proposed the so-called FCM clustering for symbolic data. They connected fuzzy clustering to deal with symbolic data.

Fuzzy data are another type like imprecise or with a source of uncertainty not caused by randomness such as linguistic assessments. This fuzzy data type is easily found in natural language, social science, knowledge representation etc. Fuzzy numbers are used to model the fuzziness of data and usually used to represent fuzzy data. Hathaway et al. [8] and Yang and Ko [12] proposed fuzzy clustering algorithms for these fuzzy data.

In real situations, we may have a date set with mixed symbolic and fuzzy data feature types. However, there is no clustering algorithm to deal with this mixed type of data. In this paper, we shall consider the feature vectors including numeric, symbolic and fuzzy data. We first modify Gowda and Diday’s dissimilarity measure for symbolic data and also change Hathaway’s parametric approach for fuzzy data. We then create a FCM clustering algorithm for these mixed feature type of data. Section 2 defines the dissimilarity measure for the mixed feature vectors. In Section 3, we present FCM clustering algorithm for these mixed data. Section 4 gives some numerical examples and comparisons. We also give a real example. Finally, conclusions are stated in Section 5.

2. Mixed type of data and its dissimilarity measure

In this section, we consider the mixed feature type of symbolic and fuzzy data. We then define its dissimilarity measure. For symbolic data components we compose the dissimilarity on the basis of the modified Gowda and Diday’s dissimilarity measure [5]. We composed part of the fuzzy data components using Hathaway’s parametric approach [8] and Yang’s dissimilarity method [12].

Suppose that any feature vector \(F\) can be written as a \(d\)-tuple of feature components \(F_1, \ldots, F_d\) with

\[
F = F_1 \times \cdots \times F_d.
\]
For any two feature vectors \( A \) and \( B \) with \( A = A_1 \times \cdots \times A_d \) and \( B = B_1 \times \cdots \times B_d \), the dissimilarity between \( A \) and \( B \) is defined as

\[
D(A, B) = \sum_{k=1}^{d} z_k D(A_k, B_k),
\]

where \( z_k \) represents the weight corresponding to the \( k \)th feature component and \( D(A_k, B_k) \) represents the dissimilarity of the \( k \)th feature component according to its feature type. Because there are symbolic and fuzzy feature components in a \( d \)-tuple feature vector, the dissimilarity \( D(A, B) \) will be the weighted sum of the dissimilarity of symbolic data which is defined by \( D_p(A_k, B_k) \) for the “position”, \( D_s(A_k, B_k) \) for the “span” and \( D_c(A_k, B_k) \) for the “content” and also with the dissimilarity of fuzzy data which is defined by \( D_f(A_k, B_k) \). Thus, \( D(A, B) \) can be totally combined with \( D_p(A_k, B_k) \), \( D_s(A_k, B_k) \), \( D_c(A_k, B_k) \) and \( D_f(A_k, B_k) \).

2.1. Symbolic feature components

Various definitions and descriptions of symbolic objects were given by Diday [3]. According to Gowda and Diday [5], the symbolic features can be divided into quantitative features and qualitative features in which each feature can be defined by \( D_p(A_k, B_k) \) due to position \( p \), \( D_s(A_k, B_k) \) due to span \( s \) and \( D_c(A_k, B_k) \) due to content \( c \). Now we review Gowda and Diday’s dissimilarity and then modify it.

(a) Quantitative type of \( A_k \) and \( B_k \): The dissimilarity between two feature components of quantitative type is defined as the dissimilarity of these values due to position, span and content.

Let \( a_l \) = lower limit of \( A_k \),

\[ a_{\mu} = \text{upper limit of } A_k, \]

\[ b_l = \text{lower limit of } B_k, \]

\[ b_{\mu} = \text{upper limit of } B_k, \]

\( \text{inters} = \text{length of intersection of } A_k \text{ and } B_k, \)

\( l_s = \text{span length of } A_k \text{ and } B_k = |\max(a_{\mu}, b_{\mu}) - \min(a_l, b_l)|, \)

\( U_k = \text{the difference between the highest and lowest values of the } k \text{th feature over all objects}, \)

\( l_a = |a_{\mu} - a_l|, \)

\( l_b = |b_{\mu} - b_l|. \)

The three dissimilarity components are then defined as follows:

\[
D_p(A_k, B_k) = \frac{|a_l - b_l|}{U_k},
\]

\[
D_s(A_k, B_k) = \frac{|l_a - l_b|}{l_s},
\]

\[
D_c(A_k, B_k) = \frac{|l_a + l_b - 2 \cdot \text{inters}|}{l_s}.
\]

Thus, \( D(A_k, B_k) = D_p(A_k, B_k) + D_s(A_k, B_k) + D_c(A_k, B_k) \). However, Gowda and Diday’s dissimilarity \( D(A_k, B_k) \) may give some bad results. In order to modify it, we compare it with the well-known
Hausdorff metric. Let \( U \) and \( V \) be any subset of a metric space \( Z \). The Hausdorff distance \( H \) on \( U \) and \( V \) is defined as

\[
H(U,V) = \max \left\{ \sup_{\mu \in U} \inf_{v \in V} d(\mu,v), \sup_{v \in V} \inf_{\mu \in U} d(\mu,v) \right\},
\]

where \( d \) is any defined metric in the metric space \( Z \). Let us consider the real space \( \mathbb{R} \). For any two intervals \( A = [a_1, a_2] \) and \( B = [b_1, b_2] \), the Hausdorff distance \( H(A,B) \) is defined with \( H(A,B) = \max\{|a_1 - b_1|, |a_2 - b_2|\} \). Table 1 gives six data sets of two objects \( A \) and \( B \) with different feature value types and their corresponding Hausdorff distance \( h_{1-14} \), Gowda and Diday’s dissimilarity \( g_1-g_{14} \) and our proposed modified dissimilarity \( md_{1-14} \).

According to the results in Table 1, we find that \( h_1 > h_2 \), \( h_3 = h_4 \), \( h_5 = h_6 \), \( h_9 > h_8 > h_7 \) and \( h_{12} > h_{11} > h_{10} \), but \( g_2 > g_1 \), \( g_4 > g_3 \), \( g_6 > g_5 \), \( g_7 = g_6 > g_8 \) and \( g_{12} > g_{10} > g_{11} \). It is clear that the results from Gowda and Diday are not good. We see that Gowda and Diday [5] defined the dissimilarity of symbolic data with quantitative features as \( D_p(A_k,B_k) \) due to position, \( D_s(A_k,B_k) \) due to span, and \( D_c(A_k,B_k) \) due to content. We find two problems with their definition of dissimilarity. The first one is about the value of \( D_p(A_k,B_k) \) due to position. If the two objects \( A \) and \( B \) have the feature types with points vs. intervals or intervals vs. intervals, then the measure of \( D_p(A_k,B_k) \) due to position should depend on the length of intervals. However, Gowda and Diday’s measure of \( D_p(A_k,B_k) \) due to position used only lower limits \( al \) and \( bl \). They did not consider upper limits \( au \) and \( bu \). The second problem is their measures of \( D_s(A_k,B_k) \) and \( D_c(A_k,B_k) \) with the division of \( ls \). However, the quantity \( ls \) should increase when the gap between two features \( A_k \) and \( B_k \) increases so that the values of \( D_s(A_k,B_k) \) and \( D_c(A_k,B_k) \) become decreasing. This situation should get bad results for the measure of dissimilarity. According to these drawbacks of Gowda and Diday’s dissimilarity [5], we propose a modified way. We find out if we replace \(|al - bl| \) with \( |(al + au)/2 - (bl + bu)/2| \) in \( D_p \) and \( ls \) with \((la + lb - inters) \) in \( D_s \) and \( D_c \), the situation will become better. In a sense, we define the modified dissimilarity by considering the middle points of intervals as the position and using the relative size of length difference for the span \( d_s(A_k,B_k) \) and the relative size of non-overlapping length for the content \( d_c(A_k,B_k) \). To standardize the scale in different features, we add the quantity \( U_k \) to the denominator of \( d_p(A_k,B_k) \), \( d_s(A_k,B_k) \) and \( d_c(A_k,B_k) \). Thus, the dissimilarity for quantitative type of \( A_k \) and \( B_k \) is modified as follows:

- due to position: \( d_p(A_k,B_k) = \frac{|(al + au)/2 - (bl + bu)/2|}{U_k} \),
- due to span : \( d_s(A_k,B_k) = \frac{|la - lb|}{U_k + la + lb - inters} \),
- due to content : \( d_c(A_k,B_k) = \frac{|la + lb - 2 \cdot inters|}{U_k + la + lb - inters} \),

and \( d(A_k,B_k) = d_p(A_k,B_k) + d_s(A_k,B_k) + d_c(A_k,B_k) \). The results with \( md_{1-14} \) using the above-modified dissimilarity are shown in Table 1. We find that the conclusions with the modified dissimilarity are exactly the same as the Hausdorff distance for the first five data sets in Table 1. These results show that our modified dissimilarity actually corrects the problems from Gowda and Diday’s dissimilarity [5]. On the other hand, we see that the dissimilarity between intervals \([0,9] \) and \([0,10] \) shall be less than the dissimilarity between \([0,1] \) and \([0,2] \) for the sixth data set in Table 1.
Table 1

Six sets of objects $A$ and $B$ with different feature types and defined distances

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>Hausdorff distance</th>
<th>Gowda and Diday dissimilarity</th>
<th>Modified dissimilarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>[5, 8]</td>
<td>$h_1 = \text{Max}{8 - 3,</td>
<td>5 - 3</td>
<td>}$</td>
</tr>
<tr>
<td>6</td>
<td>[5, 8]</td>
<td>$h_2 = \text{Max}{8 - 6,</td>
<td>5 - 6</td>
<td>}$</td>
</tr>
<tr>
<td>1</td>
<td>[1, 2]</td>
<td>$h_3 = \text{Max}{2 - 1,</td>
<td>1 - 1</td>
<td>}$</td>
</tr>
<tr>
<td>2</td>
<td>[1, 2]</td>
<td>$h_4 = \text{Max}{2 - 2,</td>
<td>1 - 2</td>
<td>}$</td>
</tr>
<tr>
<td>[2, 3]</td>
<td>[3, 6]</td>
<td>$h_5 = \text{Max}{3 - 2,</td>
<td>6 - 3</td>
<td>}$</td>
</tr>
<tr>
<td>[6, 7]</td>
<td>[3, 6]</td>
<td>$h_6 = \text{Max}{3 - 6,</td>
<td>6 - 6</td>
<td>}$</td>
</tr>
<tr>
<td>[1, 2]</td>
<td>[3, 5]</td>
<td>$h_7 = \text{Max}{3 - 1,</td>
<td>5 - 2</td>
<td>}$</td>
</tr>
<tr>
<td>[1, 2]</td>
<td>[50, 52]</td>
<td>$h_8 = \text{Max}{50 - 1,</td>
<td>52 - 2</td>
<td>}$</td>
</tr>
<tr>
<td>101</td>
<td></td>
<td>$h_9 = \text{Max}{101 - 1,</td>
<td>101 - 2</td>
<td>}$</td>
</tr>
<tr>
<td>[53, 56]</td>
<td>[57, 63]</td>
<td>$h_{10} = \text{Max}{63 - 56,</td>
<td>57 - 53</td>
<td>}$</td>
</tr>
<tr>
<td>[53, 56]</td>
<td>[64, 73]</td>
<td>$h_{11} = \text{Max}{73 - 56,</td>
<td>64 - 53</td>
<td>}$</td>
</tr>
<tr>
<td>[64, 73]</td>
<td>93</td>
<td>$h_{12} = \text{Max}{93 - 73,</td>
<td>93 - 64</td>
<td>}$</td>
</tr>
<tr>
<td>[0, 9]</td>
<td>[0, 10]</td>
<td>$h_{13} = \text{Max}{10 - 0,</td>
<td>10 - 9</td>
<td>}$</td>
</tr>
<tr>
<td>[0, 1]</td>
<td>[0, 2]</td>
<td>$h_{14} = \text{Max}{0 - 0,</td>
<td>3 - 2</td>
<td>}$</td>
</tr>
</tbody>
</table>
However, the Hausdorff distance gives a bad result with $h_{13} = h_{14} = 1$ and the modified and Gowda and Diday’s dissimilarities have good results with $md_{14} > md_{13}$ and $g_{14} > g_{13}$. We mention that the dissimilarities between two quantitative types of $A_k$ and $B_k$ for “span” and for “content” are still related to $U_k$. However, the term $U_k$ was not considered in the mathematical formula for Gowda and Diday’s $D_s(A_k, B_k)$ and $D_c(A_k, B_k)$. More comparison results will be shown and discussed in Section 4.

(b) Qualitative type of $A_k$ and $B_k$: For qualitative feature types, the dissimilarity component due to position is absent. The term $U_k$ for qualitative feature types is absent too. The two components that contribute to dissimilarity are “due to span” and “due to content”.

Let $l_a = \text{length of } A_k = \text{the number of elements in } A_k$

\[ l_b = \text{length of } B_k = \text{the number of elements in } B_k \]

\[ l_s = \text{length of } A_k \text{ and } B_k \]

\[ \text{inters} = \text{the number of elements in the intersection of } A_k \text{ and } B_k. \]

The two dissimilarity components are then defined as follows:

\[ D_s(A_k, B_k) = \frac{|l_a - l_b|}{l_s} \]

\[ D_c(A_k, B_k) = \frac{|l_a + l_b - 2 \cdot \text{inters}|}{l_s}. \]

Thus, $D(A_k, B_k) = D_s(A_k, B_k) + D_c(A_k, B_k)$. Note that the dissimilarity $D(A_k, B_k)$ for qualitative type of features are reasonable. It is not necessary to modify. That is, we have $d_s(A_k, B_k) = D_s(A_k, B_k)$ and $d_c(A_k, B_k) = D_c(A_k, B_k)$ for qualitative types of $A_k$ and $B_k$.

2.2. Fuzzy feature components

Fuzzy data types often appear in real applications. Fuzzy numbers are used to model the fuzziness of data and usually used to represent fuzzy data. The trapezoidal fuzzy numbers (TFN) are used most. Hathaway et al. [8] proposed FCM clustering for symmetric TFN using a parametric approach. They defined a dissimilarity for two symmetric TFNs and then used it for FCM clustering. However, they did not consider the left or right shapes of fuzzy numbers. It means that their dissimilarity is not suitable for LR-type fuzzy numbers. To consider FCM clustering for LR-type fuzzy numbers (including symmetric TFNs), we will provide another type of dissimilarity.

We define fuzzy data based on Hathaway’s parametric model. We extend symmetric TFNs to all TFNs by defining its parameterization as shown in Fig. 1. The notation for the parameterization of a TFN $A$ is $A = m(a_1, a_2, a_3, a_4)$ where we refer to $a_1$ as the center, $a_2$ as the inner diameter, $a_3$ as the left outer radius and $a_4$ as the right outer radius. Using this parametric representation we can parameterize the four kinds of TFNs with real numbers, intervals, triangular and trapezoidal fuzzy numbers as shown in Fig. 2.

Let $A = m(a_1, a_2, a_3, a_4)$ and $B = m(b_1, b_2, b_3, b_4)$ be any two fuzzy data. Hathaway’s distance $d_h(A, B)$ of symmetric TFNs $A$ and $B$ are defined as

\[ d_h^2(A, B) = (a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + (a_4 - b_4)^2. \]
To provide a distance which is able to be defined for all LR-type fuzzy numbers, we borrow Yang’s [12] distance definition of LR-type fuzzy numbers as follows.

Let $L$ (and $R$) be decreasing, shape functions from $\mathbb{R}^+$ to $[0,1]$ with $L(0) = 1$, $L(x) < 1$ for all $x > 0$, $L(x) > 0$ for all $x < 1$, $L(1) = 0$ or ($L(x) > 0$ for all $x$ and $L(+\infty) = 0$) (see Zimmermann [14]). A fuzzy number $X$ with its membership function $\mu_X(x)$

$$
\mu_X(x) = \begin{cases} 
L \left( \frac{m_1 - x}{\alpha} \right) & \text{for } x \leq m_1, \\
1 & \text{for } m_1 \leq x \leq m_2, \\
R \left( \frac{x - m_2}{\beta} \right) & \text{for } x \geq m_2,
\end{cases}
$$

is called an LR-type TFN. Symbolically, $X$ is denoted by $X = (m_1, m_2, \alpha, \beta)_{LR}$ where $\alpha > 0$ and $\beta > 0$ are called the left and right spreads, respectively. Given $A = (m_{1a}, m_{2a}, \alpha_a, \beta_a)_{LR}$ and $B = (m_{1b}, m_{2b}, \alpha_b, \beta_b)_{LR}$, Yang [12] defined a distance $d_{LR}(A, B)$ with

$$
d_{LR}^2(A, B) = (m_{1a} - m_{1b})^2 + (m_{2a} - m_{2b})^2 + ((m_{1a} - l\alpha_a) - (m_{1b} - l\alpha_b))^2 + ((m_{2a} + r\beta_a) - (m_{2b} + r\beta_b))^2,
$$

where $l = \int_0^1 L^{-1}(w) \, dw$ and $r = \int_0^1 R^{-1}(w) \, dw$. If $L$ and $R$ are linear, then $l = r = \frac{1}{2}$. Thus, for any given two TFNs $A = m(a_1, a_2, a_3, a_4)$ and $B = m(b_1, b_2, b_3, b_4)$, we have a distance $d_f(A, B)$ on base of Yang’s definition with

$$
d_f^2(A, B) = \left( \frac{2a_1 - a_2}{2} - \frac{2b_1 - b_2}{2} \right)^2 + \left( \frac{2a_1 + a_2}{2} - \frac{2b_1 + b_2}{2} \right)^2 + \left( \frac{2a_1 - a_2}{2} - \frac{1}{2}a_3 \right) - \left( \frac{2b_1 - b_2}{2} - \frac{1}{2}b_3 \right) \right)^2 + \left( \frac{2a_1 + a_2}{2} + \frac{1}{2}a_4 \right) - \left( \frac{2b_1 + b_2}{2} + \frac{1}{2}b_4 \right) \right)^2.
$$
Then

\[ d^2_f(A,B) = \frac{1}{4} \left( g_-^2 + g_+^2 + (g_+ - (a_3 - b_3))^2 + (g_+ + (a_4 - b_4))^2 \right), \]

where \( g_- = 2(a_1 - b_1) - (a_2 - b_2) \) and \( g_+ = 2(a_1 - b_1) + (a_2 - b_2) \).

We use the following example to show the difference between Hathaway’s \( d_h(A,B) \) and the proposed \( d_f(A,B) \). For four given constants \( a_1, a_2, a_3, a_4, \) let us consider the three triangular fuzzy numbers :\( A = m(a_1, 0, a_2, a_3), B = m(a_1, 0, a_3, a_2) \) and \( C = m(a_4, 0, a_2, a_2) \) which are shown in Fig. 3. Note that here \( a_4 = a_1 - (a_3 - a_2) \).

Thus, we have the distance \( d_h(A,B) \) and \( d_h(A,C) \) with \( d_h^2(A,B) = 2(a_2 - a_3)^2 \) and \( d_h^2(A,C) = (a_2 - a_3)^2 \). We find that \( d_h(A,B) > d_h(A,C) \). However, the proposed distances \( d_f(A,B) \) and \( d_f(A,C) \) are \( d_f^2(A,B) = \frac{1}{3}(a_2 - a_3)^2 \) and \( d_f^2(A,C) = 4(a_2 - a_3)^2 \). Thus, \( d_f(A,C) > d_f(A,B) \). The intuition tells us that the dissimilarity \( d(A,B) \) between \( A \) and \( B \) shall be less than the dissimilarity \( d(A,C) \) between \( A \) and \( C \), which means that the proposed distance \( d_f \) is much more reasonable. According to mathematical formula of Hathaway’s distances \( d_h(A,B) \) and the proposed distance \( d_f(A,B) \), we can see their difference. The main difference is that the proposed distance \( d_f(A,B) \) takes the shape functions \( L \) and \( R \) into account, but Hathaway’s distance \( d_h(A,B) \) does not, which means that our distance \( d_f(A,B) \) includes the information of shape functions \( L \) and \( R \) in the dissimilarity measures.

3. The fuzzy clustering algorithm

Let \( W = \{X_1, \ldots, X_n\} \) be a set of \( n \) feature vectors in \( \mathbb{R}^d \). Let \( c \) be a positive integer greater than one. A partition of \( W \) into \( c \) clusters can be presented using mutually disjoint sets \( W_1, \ldots, W_c \) such that \( W_1 \cup \cdots \cup W_c = W \) or equivalently by the indicator function \( \mu_1, \ldots, \mu_c \) such that \( \mu_i(X) = 1 \) if \( X \) is in \( W_i \) and \( \mu_i(X) = 0 \) if \( X \) is not in \( W_i \) for all \( i = 1, \ldots, c \). This is known as clustering \( W \) into \( c \) clusters \( W_1, \ldots, W_c \) using a so-called hard \( c \)-partition \( \{\mu_1, \ldots, \mu_c\} \). The fuzzy extension allows \( \mu_i(X) \) to be membership functions in fuzzy sets \( \mu_i \) on \( W \) assuming values in the interval \([0,1]\) such that \( \sum_{i=1}^c \mu_i(X) = 1 \) for all \( X \) in \( W \). In this case, \( \{\mu_1, \ldots, \mu_c\} \) is called a fuzzy \( c \)-partition of \( W \). Thus, the fuzzy \( c \)-mean (FCM) objective function \( J(\mu,A) \) is defined as

\[
J(\mu,A) = \sum_{i=1}^c \sum_{j=1}^n \mu_i^m(X_j) \|X_j - A_i\|^2,
\]

where \( m \) is a fixed number bigger than one to present the degree of fuzziness and \( \{\mu_1, \ldots, \mu_c\} \) is a fuzzy \( c \)-partition and \( \{A_1, \ldots, A_c\} \) is a set of cluster centers. The FCM clustering is an
An iterative algorithm through the necessary conditions for minimizing $J(\mu, A)$ with the following update equations:

$$A_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m X_j}{\sum_{j=1}^{n} \mu_{ij}^m}, \quad i = 1, \ldots, c$$

(2)

and

$$\mu_{ij} = \mu_i(X_j) = \left( \sum_{k=1}^{c} \frac{||X_j - A_i||^2/(m-1)}{||X_j - A_k||^2/(m-1)} \right)^{-1}, \quad i = 1, \ldots, c, \quad j = 1, \ldots, n.$$  

(3)

If the feature vectors are numeric data in $\mathbb{R}^d$, the FCM clustering algorithm is well used. However, in applying the FCM to symbolic objects, there are problems encountered, such as the weighted mean equation (2) and the Euclidean distance $\| \cdot \|$ not suitable for symbolic objects. To overcome these problems, El-Sonbaty and Ismail [4] proposed a new representation way for cluster centers.

A cluster center is assumed to be formed as a group of features and each feature is composed of several events. Let $A_{k|p|}$ be the $p$th event of feature $k$ in cluster $i$ and let $e_{k|p|i}$ be the membership degree of association of the $p$th event $A_{k|p|}$ to the feature $k$ in cluster $i$. Thus, the $k$th feature of the $i$th cluster center $A_{ik}$ can be presented as:

$$A_{ik} = [(A_{k|1|i}, e_{k|1|i}), \ldots, (A_{k|p|i}, e_{k|p|i})].$$  

(4)

In this case, we shall have

$$0 \leq e_{k|p|i} \leq 1 \quad \text{and} \quad \sum_p e_{k|p|i} = 1,$$

$$\bigcap_p A_{k|p|i} = \phi \quad \text{and} \quad \bigcup_p A_{k|p|i} = \bigcup_j X_{jk},$$

where $e_{k|p|i} = 0$ if the event $A_{k|p|i}$ is not a part of feature $k$ in the cluster center $A_i$ and $e_{k|p|i} = 1$ if there are no other events than the event $A_{k|p|i}$ sharing this event in forming the feature $k$ in cluster center $A_i$. Thus, the update equation for $e_{k|p|i}$ is

$$e_{k|p|i} = \frac{\sum_{j=1}^{n} \mu_{ij}^m \theta}{\sum_{j=1}^{n} \mu_{ij}^m},$$  

(5)

where $\theta \in \{0, 1\}$ and $\theta = 1$ if the $k$th feature of the $j$th datum $X_j$ consists of the $p$th event, otherwise $\theta = 0$. $\mu_{ij} = \mu_i(X_j)$ is the membership of $X_j$ in cluster $i$.

The membership function $e_{k|p|i}$ is an important index function proposed by El-Sonbaty and Ismail [4] for using the FCM with symbolic data. In this paper, except applying the FCM to symbolic feature components, we also consider the FCM to fuzzy feature components. Based on the parametric model $A = m(a_1, a_2, a_3, a_4)$ for the fuzzy datum $A$, the parameter $(a_1, a_2, a_3, a_4)$ is in $\mathbb{R}^d$. Thus, each fuzzy feature is composed of the event \{a_1, a_2, a_3, a_4\}. Let $A_{ik}$ be the $k$th fuzzy feature component of the $i$th cluster center with the parametric form $A_{ik} = m(a_{ik1}, a_{ik2}, a_{ik3}, a_{ik4})$. Because the parametric vector $(a_{ik1}, a_{ik2}, a_{ik3}, a_{ik4})$ is a numeric vector data, we shall set up an indicator function $e_{k|i}$ to indicate this fuzzy feature component $A_{ik}$ where $e_{k|i} = 1$ if the $k$th feature in the cluster $i$ is the fuzzy feature.
$A_{ik}$, otherwise $e_k|_{i} = 0$. Next, we set up the FCM objective function for these mixed features of symbolic and fuzzy data.

Let \{X_1, \ldots, X_n\} be a data set of mixed feature types. The FCM objective function is defined as

$$J(\mu, e, a) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_i^m d^2(X_j, A_i),$$

where

$$d^2(X_j, A_i) = \sum_{k'}^{\text{of symbolic}} \left( \sum_{p} d^2(X_{jk'}, A_{k'p|_{i}}) \cdot e_{k'p|_{i}} \right) + \sum_{k}^{\text{of fuzzy}} d^2_{f}(X_{jk}, A_{ik}).$$

There are parameters $\{\mu_1, \ldots, \mu_c\}$, $\{e_{k'p|_{i}}\}$ and $\{a_{ik1}, a_{ik2}, a_{ik3}, a_{ik4}\}$ for the FCM objective function $J(\mu, e, a)$. The Picard’s method for approximating the optimal solutions which minimize the objective function $J(\mu, e, a)$ is considered with the necessary condition of a minimizer of $J$ over the parameter $\mu$. Consider the Lagrangian

$$L(\mu, \lambda) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_i^m d^2(X_j, A_i) - \lambda \left( \sum_{i=1}^{c} \mu_i - 1 \right).$$

Take the derivative of $L$ w.r.t. $\mu_{ij}$ and $\lambda$, we find that

$$\mu_{ij} = \left( \sum_{q=1}^{c} \frac{(d^2(X_j, A_i))^{1/(m-1)} \lambda}{(d^2(X_j, A_q))^{1/(m-1)}} \right)^{-1}, \quad i = 1, \ldots, c, \ j = 1, \ldots, n,$$

where $d^2(X_j, A_i)$ is defined by Eq. (7) for which $d^2(X_{jk'}, A_{k'p|_{i}})$ and $d^2_{f}(X_{jk}, A_{ik})$ are the dissimilarities for symbolic and fuzzy data proposed in Section 2.

We then consider the other two groups of parameters $e$ and $a$. It is known that the parameters $e_{k'p|_{i}}$ are for these $k'$ which are symbolic and the parameters $\{a_{ik1}, a_{ik2}, a_{ik3}, a_{ik4}\}$ are for these $k$ which are fuzzy.

(a) For these $k'$ which are symbolic, taking the derivative of $J(\mu, e, a)$ over $e$ is equivalent to taking the derivative of $J(\mu, e)$ with

$$J(\mu, e) = \sum_{i=1}^{c} \sum_{j=1}^{n} \sum_{k'}^{\text{of symbolic}} \left( \sum_{p} d^2(X_{jk'}, A_{k'p|_{i}}) \cdot e_{k'p|_{i}} \right).$$

Based on the similar procedure of El-Sonbaty and Ismail [4], we have the following update equation for $e_{k'p|_{i}}$ as

$$e_{k'p|_{i}} = \frac{\sum_{j=1}^{n} \mu_{ij}^m \cdot \theta}{\sum_{j=1}^{n} \mu_{ij}^m}.$$
(b) For these \( k \) which are fuzzy, taking the derivative of \( J(\mu, e, a) \) over \( a \) is equivalent to taking the derivative of \( J(\mu, a) \) with

\[
J(\mu, a) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} \sum_{k \text{ of fuzzy}} d_{f}^{2}(X_{jk}, A_{ik}),
\]

where \( X_{jk} = m(x_{jk1}, x_{jk2}, x_{jk3}, x_{jk4}) \), \( A_{ik} = m(a_{ik1}, a_{ik2}, a_{ik3}, a_{ik4}) \) and

\[
d_{f}^{2}(X_{jk}, A_{ik}) = \left( \frac{1}{4} \right) \left( (2(x_{jk1} - a_{ik1}) - (x_{jk2} - a_{ik2}))^2 + (2(x_{jk1} - a_{ik1}) + (x_{jk2} - a_{ik2}))^2 
+ (2(x_{jk1} - a_{ik1}) - (x_{jk2} - a_{ik2}) - (x_{jk3} - a_{ik3}))^2 
+ (2(x_{jk1} - a_{ik1}) + (x_{jk2} - a_{ik2}) + (x_{jk4} - a_{ik4}))^2 \right).
\]

Taking the partial derivative of \( J(\mu, a) \) over \( a \) and setting to zero, we have

\[
\frac{\partial J}{\partial a_{ik1}} = \sum_{j=1}^{n} ( -1 ) \mu_{ij}^{m} \left( 8(x_{jk1} - x_{jk3} - 8a_{ik1} + a_{ik3} - a_{ik4}) \right) = 0,
\]

\[
\frac{\partial J}{\partial a_{ik2}} = \sum_{j=1}^{n} \left( - \frac{1}{2} \right) \mu_{ij}^{m} \left( 4x_{jk2} + x_{jk3} + x_{jk4} - 4a_{ik2} - a_{ik3} - a_{ik4} \right) = 0,
\]

\[
\frac{\partial J}{\partial a_{ik3}} = \sum_{j=1}^{n} \left( \frac{1}{2} \right) \mu_{ij}^{m} \left( 2x_{jk1} - x_{jk2} - x_{jk3} - 2a_{ik1} + a_{ik2} + a_{ik3} \right) = 0,
\]

\[
\frac{\partial J}{\partial a_{ik4}} = \sum_{j=1}^{n} \left( - \frac{1}{2} \right) \mu_{ij}^{m} \left( 2x_{jk1} + x_{jk2} + x_{jk4} - 2a_{ik1} - a_{ik2} - a_{ik4} \right) = 0.
\]

Thus, we have the update equations as follows.

\[
a_{ik1} = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} (8x_{jk1} - x_{jk3} + x_{jk4} + a_{ik3} - a_{ik4})}{8 \sum_{j=1}^{n} \mu_{ij}^{m}}, \quad (10)
\]

\[
a_{ik2} = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} (4x_{jk2} + x_{jk3} + x_{jk4} - a_{ik3} - a_{ik4})}{4 \sum_{j=1}^{n} \mu_{ij}^{m}}, \quad (11)
\]

\[
a_{ik3} = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} (-2x_{jk1} + x_{jk2} + x_{jk3} + 2a_{ik1} - a_{ik2})}{\sum_{j=1}^{n} \mu_{ij}^{m}}, \quad (12)
\]

\[
a_{ik4} = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} (2x_{jk1} + x_{jk2} + x_{jk4} - 2a_{ik1} - a_{ik2})}{\sum_{j=1}^{n} \mu_{ij}^{m}}. \quad (13)
\]

On the basis of these necessary conditions we can construct an iterative algorithm.
Table 2
Memberships for symbolic data with FSCM and proposed MVFCM

<table>
<thead>
<tr>
<th>Data</th>
<th>FSCM $\mu_{ij}$</th>
<th>MVFCM $\mu_{ij}$</th>
<th>FSCM $\mu_{ij}$</th>
<th>MVFCM $\mu_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8435</td>
<td>0.1565</td>
<td>0.9929</td>
<td>0.0071</td>
</tr>
<tr>
<td>3</td>
<td>0.8834</td>
<td>0.1166</td>
<td>0.9940</td>
<td>0.0060</td>
</tr>
<tr>
<td>4</td>
<td>0.8391</td>
<td>0.1609</td>
<td>0.9946</td>
<td>0.0054</td>
</tr>
<tr>
<td>[3, 5]</td>
<td>0.3923</td>
<td>0.6077</td>
<td>0.9925</td>
<td>0.0075</td>
</tr>
<tr>
<td>[6, 8]</td>
<td>0.5585</td>
<td>0.4415</td>
<td>0.9785</td>
<td>0.0215</td>
</tr>
<tr>
<td>15</td>
<td>0.9354</td>
<td>0.0646</td>
<td>0.1409</td>
<td>0.8591</td>
</tr>
<tr>
<td>16</td>
<td>0.9497</td>
<td>0.0503</td>
<td>0.0803</td>
<td>0.9197</td>
</tr>
<tr>
<td>[16, 20]</td>
<td>0.2395</td>
<td>0.7605</td>
<td>0.0463</td>
<td>0.9537</td>
</tr>
<tr>
<td>17</td>
<td>0.9224</td>
<td>0.0776</td>
<td>0.0439</td>
<td>0.9561</td>
</tr>
<tr>
<td>29</td>
<td>0.7107</td>
<td>0.2893</td>
<td>0.0477</td>
<td>0.9523</td>
</tr>
</tbody>
</table>

3.1. Mixed-type variables FCM (MVFCM)

Step 1: Fix $m$ and $c$. Given an $\varepsilon > 0$. Initialize a fuzzy $c$-partition $\mu^{(0)} = \{\mu_1^{(0)}, \ldots, \mu_c^{(0)}\}$. Set $\ell = 0$.

Step 2: For symbolic feature $k'$, compute $i$th cluster center $A_{k'}^{(\ell)} = [(A_{k'1}^{(\ell)}; e_{k'1}^{(\ell)}), \ldots, (A_{k'p}^{(\ell)}; e_{k'p}^{(\ell)})]$ using Eq. (9). For fuzzy feature $k$, compute $i$th cluster center $A_k^{(\ell)} = (a_{ik1}; a_{ik2}; a_{ik3}; a_{ik4})$ using Eqs. (10)–(13).

Step 3: Update to $\mu^{(\ell+1)}$ using Eqs. (7) and (8).

Step 4: Compare $\mu^{(\ell+1)}$ to $\mu^{(\ell)}$ in a convenient matrix norm.

IF $\|\mu^{(\ell+1)} - \mu^{(\ell)}\| < \varepsilon$, THEN STOP.
ELSE $\ell = \ell + 1$ and GOTO Step 2.

4. Experimental results and comparisons

In this section, the proposed modified dissimilarity measure and algorithms for mixed feature variables are used in three examples. The comparisons are also made. We use two artificial data in Examples 1 and 2 and then apply it to a real data in Example 3.

Example 1. Consider a data set of size 10 with \{2, 3, 4, [3, 5], [6, 8], 15, 16, [16, 20], 17, 29\}. We implement the two algorithms with fuzzy symbolic $c$-means algorithm (FSCM) of El-Sonbaty and Ismail [4] and our mixed-type variable FCM (MVFCM) for this data set. The membership results are shown in Table 2.

According to the results in Table 2, we find that FSCM divides two clusters with $C_1 = \{[3, 5], [16, 20]\}$ and $C_2 = \{2, 3, 4, [6, 8], 15, 16, 17, 29\}$. However, our MVFCM algorithm divides two cluster with $C_1 = \{2, 3, 4, [3, 5], [6, 8]\}$ and $C_2 = \{15, 16, [16, 20], 17, 29\}$. It is clear that our MVFCM algorithm gives better results.
Table 3
Memberships for fuzzy data with Hathaway’s FCM and MVFCM

<table>
<thead>
<tr>
<th>Data</th>
<th>Hathaway’s FCM</th>
<th>MVFCM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu_{1j}$</td>
<td>$\mu_{2j}$</td>
</tr>
<tr>
<td>$A = [0, 0, 1, 1]$</td>
<td>0.9666</td>
<td>0.0334</td>
</tr>
<tr>
<td>$B = [0, 0, 3, 3]$</td>
<td>0.6620</td>
<td>0.3380</td>
</tr>
<tr>
<td>$C = [0, 0, 5, 5]$</td>
<td>0.1744</td>
<td>0.8256</td>
</tr>
<tr>
<td>$D = [2, 0, 3, 3]$</td>
<td>0.3153</td>
<td>0.6847</td>
</tr>
<tr>
<td>$E = [4, 0, 5, 5]$</td>
<td>0.1234</td>
<td>0.8766</td>
</tr>
</tbody>
</table>

Example 2. We consider a data set with five triangular fuzzy number $A = [0, 0, 1, 1]$, $B = [0, 0, 3, 3]$, $C = [0, 0, 5, 5]$, $D = [2, 0, 3, 3]$ and $E = [4, 0, 5, 5]$. These fuzzy numbers are shown in Fig. 4. We implement the Hathaway’s FCM and the proposed MVFCM to this data set. The membership results are shown in Table 3.

According to the results in Table 3, the Hathaway’s FCM gives two clusters with $C_1 = \{A, B\}$ and $C_2 = \{C, D, E\}$. However, our MVFCM gives $C_1 = \{A, B, C\}$ and $C_2 = \{D, E\}$. Intuitively, MVFCM gives more reasonable results than Hathaway’s FCM.

Example 3. In this example, we use a real data. There are 10 brands of automobiles from four companies Ford, Toyota, China-Motor and Yulon-Motor in Taiwan. The data set is shown in Table 4.

In each brand, there are six feature components—company, exhaust, price, color, comfort and safety features. In the color feature, the notations W = white, S = silver, D = dark, R = red, B = blue, G = green, P = purple, Gr = grey and Go = golden are used. In all feature components, company, exhaust, color are symbolic data and price are real data and comfort and safety are fuzzy data. We used the dissimilarity measure described in Section 2 to illustrate the dissimilarity calculation between object one of Virage and the object five of M2000 as follows:

$$D(\text{Virage, M2000}) = D(\text{China-Motor, Ford}) + D(1.8, 2.0)$$
$$+ D(63.9, 64.6) + D(\{W, S, D, R, B\}, \{W, S, D, G, Go\})$$
$$+ D([10, 0, 2, 2],[8, 0, 2, 2]) + D([9, 0, 3, 3],[9, 0, 3, 3]),$$
Table 4  
Data set of automobiles

<table>
<thead>
<tr>
<th>No.</th>
<th>Brands</th>
<th>Company</th>
<th>Exhaust(L)</th>
<th>Price (NT$10000)</th>
<th>Color</th>
<th>Comfort</th>
<th>Safetiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Virage</td>
<td>China-Motor</td>
<td>1.8</td>
<td>63.9</td>
<td>W, S, D, R, B</td>
<td>[10,0,2,2]</td>
<td>[9,0,3,3]</td>
</tr>
<tr>
<td>2</td>
<td>New Lancer</td>
<td>China-Motor</td>
<td>1.8</td>
<td>51.9</td>
<td>W, S, D, R, G</td>
<td>[6,0,2,2]</td>
<td>[6,0,3,3]</td>
</tr>
<tr>
<td>3</td>
<td>Galant</td>
<td>China-Motor</td>
<td>2.0</td>
<td>71.8</td>
<td>W, S, R, G, P, Gr</td>
<td>[12,4,2,0]</td>
<td>[15,5,3,0]</td>
</tr>
<tr>
<td>4</td>
<td>Tierra Activa</td>
<td>Ford</td>
<td>1.6</td>
<td>46.9</td>
<td>W, S, D, R, G, Go</td>
<td>[6,0,2,2]</td>
<td>[6,0,3,3]</td>
</tr>
<tr>
<td>5</td>
<td>M2000</td>
<td>Ford</td>
<td>2.0</td>
<td>64.6</td>
<td>W, S, D, G, Go</td>
<td>[8,0,2,2]</td>
<td>[9,0,3,3]</td>
</tr>
<tr>
<td>6</td>
<td>Tercel</td>
<td>Toyota</td>
<td>1.5</td>
<td>45.8</td>
<td>W, S, R, G</td>
<td>[4,4,0,2]</td>
<td>[6,0,3,3]</td>
</tr>
<tr>
<td>7</td>
<td>Corolla</td>
<td>Toyota</td>
<td>1.8</td>
<td>74.3</td>
<td>W, S, D, R, G</td>
<td>[12,4,2,0]</td>
<td>[12,0,3,3]</td>
</tr>
<tr>
<td>8</td>
<td>Premio G2.0</td>
<td>Toyota</td>
<td>2.0</td>
<td>72.9</td>
<td>W, S, D, G</td>
<td>[10,0,2,2]</td>
<td>[15,5,3,0]</td>
</tr>
<tr>
<td>9</td>
<td>Cerfiro</td>
<td>Yulon-Motor</td>
<td>2.0</td>
<td>69.9</td>
<td>W, S, D</td>
<td>[8,0,2,2]</td>
<td>[12,0,3,3]</td>
</tr>
<tr>
<td>10</td>
<td>March</td>
<td>Yulon-Motor</td>
<td>1.3</td>
<td>39.9</td>
<td>W, R, G, P</td>
<td>[4,4,0,2]</td>
<td>[3,5,0,3]</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
D(\text{China} − \text{Motor, Ford}) &= (|1 − 1|/2)^2 + (|1 + 1 − 2 \times 0|/2)^2 = 1, \\
D(1.8, 2.0) &= [(1.8 + 1.8)/2 − (2.0 + 2.0)/2]/(2.0 − 1.3))^2 \\
&+[(0.0 − 0)/(0.7 + 0 + 0 − 0)]^2 + [(0 + 0 − 0.2 \times 0)/(0.7 + 0 + 0 − 0)]^2 \\
&= 0.0816, \\
D(63.9, 64.6) &= \{[2 \times (63.9 − 64.6)]^2 + [2 \times (63.9 − 64.6)]^2 \\
&+ [2 \times (63.9 − 64.6)]^2 + [2 \times (63.9 − 64.6)]^2\}/4 = 1.96, \\
D(\{W, S, D, R, B\}, \{W, S, D, G, Go\}) &= |5 − 5|/(5 + 5 − 3))^2 + |5 + 5 − 2 \times 3|/(5 + 5 − 3)^2 = 0.3265, \\
D([10,0,2,2],[8,0,2,2]) &= \{[2 \times (10 − 8) − (0 − 0)]^2 + [2 \times (10 − 8) + (0 − 0)]^2 \\
&+[(2 \times (10 − 8) − (0 − 0)) − (2 − 2)]^2 + [(2 \times (10 − 8) + (0 − 0)) + (2 − 2)]^2\}/4 = 16.
\end{align*}
\]

Similarly,

\[
D([9,0,3,3],[9,0,3,3]) = 0.
\]

Thus,

\[
D(\text{Virage, M2000}) = 1 + 0.0816 + 1.96 + 0.3265 + 16 + 0 = 19.3681.
\]

In order to illustrate the structure of a cluster center for symbolic data, we give memberships of these 10 data points as shown in Table 5.
Table 5
Memberships of data points for a cluster center

<table>
<thead>
<tr>
<th>No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu_{ij})</td>
<td>0.4</td>
<td>0.3</td>
<td>0.35</td>
<td>0.5</td>
<td>0.25</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 6
Structure of symbolic feature components for a cluster center

<table>
<thead>
<tr>
<th>Company</th>
<th>Exhaust</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>China-Motor</td>
<td>0.3182</td>
<td>W, S, D, R, B</td>
</tr>
<tr>
<td>Ford</td>
<td>0.2273</td>
<td>W, S, D, R, G</td>
</tr>
<tr>
<td>Toyota</td>
<td>0.2424</td>
<td>W, S, R, G, Go</td>
</tr>
<tr>
<td>Yulon-Motor</td>
<td>0.2121</td>
<td>W, S, D, G, Gr</td>
</tr>
<tr>
<td></td>
<td>1.8</td>
<td>0.2121</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.3333</td>
</tr>
<tr>
<td></td>
<td>1.6</td>
<td>0.1515</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.2424</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>0.0606</td>
</tr>
<tr>
<td>W, S, D, R, G</td>
<td>0.0909</td>
<td></td>
</tr>
<tr>
<td>W, S, D, R, G, G</td>
<td>0.1061</td>
<td></td>
</tr>
<tr>
<td>W, S, D, R, G, Go</td>
<td>0.1515</td>
<td></td>
</tr>
<tr>
<td>W, S, R, G</td>
<td>0.2424</td>
<td></td>
</tr>
<tr>
<td>W, S, D, G</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>W, S, D, G</td>
<td>0.1515</td>
<td></td>
</tr>
<tr>
<td>W, R, G, P</td>
<td>0.0606</td>
<td></td>
</tr>
</tbody>
</table>

We find out the structure of a cluster center for symbolic feature components. For the cluster center, we have its membership with \(0.4 + 0.3 + 0.35 + 0.5 + 0.25 + 0.8 + 0 + 0 + 0.5 + 0.2 = 3.3\). Thus, for the symbolic feature of company, we have the memberships of the cluster center with

- China-Motor: \((0.4 + 0.3 + 0.35)/3.3 = 0.3182\),
- Ford: \((0.5 + 0.25)/3.3 = 0.2273\),
- Toyota: \((0.8 + 0 + 0)/3.3 = 0.2424\),
- Yulon-Motor: \((0.5 + 0.2)/3.3 = 0.2121\).

Similarly, we can find memberships of other symbolic feature components for the cluster center as shown in Table 6.

Now we implement the proposed MVFCM algorithm for this mixed variables of auto data set where we choose \(m = 2\), \(c = 2\) and \(\varepsilon = 0.0001\). The results of memberships of 10 data points are shown in Table 7. According results in Table 7, we have two clusters with \(C_1 = \{\text{Virage, New Lancer, Galant, M2000, Corolla, Premio G2.0, Cerfiro}\}\) and \(C_2 = \{\text{Tierra Activa, Tercel, March}\}\). Intuitively, the results are very reasonable.

Finally, we may be concerned of algorithms about the sensitivity to initializations. On the basis of more numerical experiments, we find that both of El-Sonbaty and Ismail’s fuzzy symbolic \(c\)-means (FSCM) and our MVFCM are similar to the FCM that are all sensitive to initializations. However, the sensitivity to initials are all dependent to the structure of data. If clusters in data are separated well, then these algorithms are not sensitive to initials. If the degree of overlapping between clusters in the data grows, then the sensitivity of these algorithms to initials will increase.
Table 7
Results for auto data with the proposed MVFCM

<table>
<thead>
<tr>
<th>No.</th>
<th>(\mu_1)</th>
<th>(\mu_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9633</td>
<td>0.0367</td>
</tr>
<tr>
<td>2</td>
<td>0.9633</td>
<td>0.0367</td>
</tr>
<tr>
<td>3</td>
<td>0.9951</td>
<td>0.0049</td>
</tr>
<tr>
<td>4</td>
<td>0.0966</td>
<td>0.9034</td>
</tr>
<tr>
<td>5</td>
<td>0.9951</td>
<td>0.0049</td>
</tr>
<tr>
<td>6</td>
<td>0.0135</td>
<td>0.9865</td>
</tr>
<tr>
<td>7</td>
<td>0.9633</td>
<td>0.0367</td>
</tr>
<tr>
<td>8</td>
<td>0.9951</td>
<td>0.0049</td>
</tr>
<tr>
<td>9</td>
<td>0.9951</td>
<td>0.0049</td>
</tr>
<tr>
<td>10</td>
<td>0.0185</td>
<td>0.9815</td>
</tr>
</tbody>
</table>

5. Conclusions

A new fuzzy clustering algorithm called a mixed-type variable fuzzy c-means (MVFCM) was proposed to deal with mixed feature variables. We defined the dissimilarity measure for these mixed features of data and then created the algorithm. Most of the clustering algorithms can only treat the same type of data features. The proposed MVFCM clustering algorithm allows different types of data features such as numeric, symbolic and fuzzy data.

The experimental results demonstrated that the MVFCM algorithm is effective in treating mixed feature variables. It can produce the fuzzy partitions and present cluster center structures for mixed feature data. In real situations, the real data are often a mixed type of numeric, symbolic and fuzzy features. In these cases, the proposed MVFCM algorithm should be useful and effective as a data analysis tool.

Acknowledgements

The authors thank the anonymous referees for their helpful comments and suggestions to improve the presentation of the paper. This work was supported in part by the National Science Council of Taiwan, under grant NSC-89-2118-M-033-004.

References