Neurocomputing 166 (2015) 422–435

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Categorical fuzzy *k*-modes clustering with automated feature weight learning

Arkajyoti Saha [a], Swagatam Das [b,*]

[a] Stat-math Unit, Indian Statistical Institute, Kolkata 700108, India
[b] Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 26 September 2014
Received in revised form
9 January 2015
Accepted 19 March 2015
Communicated by T. Heskes
Available online 23 April 2015

Keywords:
Fuzzy clustering
WFk-modes
Fuzzy K-modes
Automated feature weights
Categorical data

ABSTRACT

This article presents and investigates a new variant of the fuzzy *k*-Modes clustering algorithm for categorical data with automated feature weight learning. The modification strengthens the classical fuzzy *k*-Modes algorithm by associating higher weights to features which are instrumental in recognizing the clustering pattern of the data. A statistical comparison between the performances of the proposed algorithm and the conventional fuzzy *k*-Modes algorithm on synthetic and real world datasets, have been carried out with respect to mean values, best performance count, and medians. We take a novel approach towards the comparison of the fuzziness of the obtained clusters. To the best of our knowledge, such comparison has been reported here for the first time for the case of categorical data. The results obtained, shows that the proposed algorithm enjoys an edge over the conventional fuzzy *k*-Modes algorithm both in terms of Rand Index and fuzziness measures.

1. Introduction

Clustering is a well-known technique in data mining to partition the given data on the basis of some similarity/dissimilarity measures. Recently clustering of the categorical data has received increasing attention, as it forms an important part of real world data analysis. The fuzzy version of the *k*-Means algorithm [1] served as the inspiration to the introduction of the fuzzy *k*-Modes algorithm for clustering categorical data [2].The conventional fuzzy *k*-Modes algorithm was extended by Kim et al. [3] by representing the clusters of categorical data with fuzzy centroids instead of the hard-type centroids used in the original algorithm. He et al. [4] and San et al. [5] independently proposed a new dissimilarity metric for the *k*-Modes clustering process in order to make the clustering results more accurate. The fundamental idea underlying their work was to use the relative attribute frequencies of the cluster modes as the similarity measure in the *k*-modes objective function. Bai and Liang [6] recently proposed a new optimization framework for the *k*-Modes algorithm for considering the between cluster information in a meaningful way. Cao et al. [7] came up with a weighted *k*-Modes algorithm for the subspace clustering of categorical data. Ng et al. [8] rigorously proved that the object cluster membership assignment method and the mode updating formulae under the new dissimilarity measure indeed minimize the objective function. Since the conventional fuzzy *k*-Modes clustering algorithm [2] may occasionally stop at locally optimal solutions, Gan et al. [9] hybridized a Genetic Algorithm (GA) with the fuzzy *k*-Modes algorithm to facilitate the global optimization of the underlying objective function. Yang et al. [10] proposed a block fuzzy *k*-Modes clustering algorithm to construct simultaneously an optimal partition of objects and also attribute variables into homogeneous blocks. Recently Bai et al. [11] improved the objective function of the fuzzy *k*-Modes algorithm by adding the between-cluster information so that the within-cluster dispersion can be minimized and the between-cluster separation can be increased. Cao et al. [12] proposed a new dissimilarity measure for the *k*-Modes algorithm based on the idea of biological and genetic taxonomy and rough membership function. Saha et al. proposed a rough set based fuzzy *k*-Modes clustering algorithm in [13]. Similar to the approach of [9], Soliman et al. [14] proposed to integrate the Particle Swarm Optimization (PSO) method with the standard fuzzy *k*-Modes clustering algorithm to facilitate the global minimization of the objective function involved.

An important problem associated with the conventional fuzzy *k*-Modes algorithm is that it treats each variable equally likely, which may deteriorate the clustering performance by putting emphasis on features which do not contribute in demarcating

* Corresponding author.
E-mail addresses: arkajyotisaha93@gmail.com (A. Saha), swagatam.das@isical.ac.in (S. Das).

http://dx.doi.org/10.1016/j.neucom.2015.03.037
0925-2312/

the clusters. The algorithm cannot appropriately down weight the redundant feature(s), therefore leading to a less accurate clustering performance. The concept of variable weighting originated from the SYNCLUS algorithm due to Desarbo et al. [15]. The concept of automated variable weighting for the hard k-means clustering was introduced by Huang et al. [16].

In this article we propose a weighted fuzzy k-Modes algorithm which can automatically detect the appropriate weights of the concerned variables, according to their contributions in clustering. Here we actually modify the weights in each iteration of the Alternative Optimization (AO) method used to minimize the objective function of the fuzzy k-Modes algorithm, based on the current partition of the data. We present a series of experimental results on performances of the new algorithm and the conventional fuzzy k-Modes algorithm [2] on standard synthetic as well as real world categorical data sets [17]. Our experimental investigations focus on perfection of the fuzzy partition obtained (in terms of Rand Index) as well as the fuzziness of the partition obtained (which is first of its kind, as far as our knowledge is concerned).

The rest of the paper is organized as follows: Section 2 consists of a brief description of the conventional fuzzy k-Modes algorithm. Section 3 provides a detailed description of the new Weighted Fuzzy k-Modes (WFk-Modes) algorithm for clustering categorical data. In Section 4 we present and discuss the performance measures and the experimental results. Finally the paper is concluded in Section 5.

2. Fuzzy k-modes algorithm for clustering categorical data

Let the set of objects to be clustered be defined by a set of d attributes $A_1, A_2, \ldots A_p$. Each attribute A_l describes a domain of values, denoted by $DOM(A_l)$. Here, we concern ourselves with the categorical data type. A domain $DOM(A_l)$ is defined as categorical if it is finite and unordered, i.e. if $a_{il}, a_{jl} \in DOM(A_l)$, either $a_{il} = a_{jl}$ or $a_{il} \neq a_{jl}$. This indicates there does not exist any order among the elements of $DOM(A_l)$. $X = [x_1, x_2, \ldots, x_p]$ is called a categorical object if x_l takes categorical values $\forall\, l = 1, 2, \ldots, p$. If $x_l \in DOM(A_l) \forall\, l = 1, 2, \ldots, p$, we can represent X as the following conjunction of attribute-value pairs:

$$X = \bigwedge_{p}^{l=1} [A_l = x_l].$$

When the value of an attribute A_l is missing, then we denote the attribute value of A_l by ϵ. Let the data under consideration be defined as $X = (X_1, X_2, \ldots, X_n)$ where X_i is the i^{th} data point (p dimensional) and $X_i = [x_{i1}, x_{i2}, \ldots, x_{il}]$.

For two data points X_s and X_t, we define $X_s = X_t$, if and only if, $x_{sl} = x_{tl} \forall\, l = 1, 2, \ldots, p$.Though the relation $X_s = X_t$ does not mean that X_s and X_t are the same object in real-world database, but rather that two objects have equal values in attributes $A_1, A_2, \ldots, A_p$.

The traditional fuzzy k-Modes algorithm was developed by Huang and Ng [2]. Assuming n objects and k clusters, the objective function to be minimized in fuzzy k-Modes algorithm can be expressed as

$$O^m(U, Z, X) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ji}^m d_{ji}, \tag{1}$$

where m is the fuzzifier value and $U = [u_{ji}]_{k \times n}$ is the membership matrix. Also in (1),

$X = (X_1, X_2, \ldots, X_n)$ is the data matrix, where X_i is the ith data point(p dimensional),

$Z = (Z_1, Z_2, \ldots, Z_k)$ i.e. the cluster centers, where Z_j is centre of the jth cluster,

$d_{ji} = d(Z_j, X_i) = \sum_{l=1}^{p} \delta(z_{jl}, x_{il})$ where $z_{jl} = l$th component of Z_j, x_{il} is the lth term of X_i.

$$\delta(z_{jl}, x_{il}) = \begin{cases} 0, & \text{if } z_{jl} = x_{il} \\ 1, & \text{if } z_{jl} \neq x_{il} \end{cases}$$

The minimization should be carried out subject to

$$\sum_{j=1}^{k} u_{ji} = 1, 1 \leq i \leq n \text{ and } 0 < \sum_{i=1}^{n} u_{ji} < n, \quad 1 \leq j \leq k$$

The steps of the conventional fuzzy k-Modes algorithm are presented below. Here $Z^{(t)}$denotes centers at t^{th} iteration and$U^{(t)}$ is the membership matrix at the t^{th} iteration.

(1) Randomly select $k$data points $Z^{(1)}$as modes. Fix the fuzzifier value (m=1.1 in our case). Determine membership matrix $U^{(1)}$ according to Eq. (2). Set iteration counter $t = 1$.
(2) Determine $Z^{(t+1)}$ according to Eq. (2). If $O^m\left(U^{(t)}, Z^{(t+1)}, X\right) = O^m\left(U^{(t)}, Z^{(t)}, X\right)$, then stop. Otherwise go to step (c).
(3) Determine $U^{(t+1)}$ according to Eq. (2). If $O^m\left(U^{(t+1)}, Z^{(t+1)}, X\right) = O^m\left(U^{(t)}, Z^{(t+1)}, X\right)$, then stop. Otherwise set $t = t+1$, go to step (b).

2.1. Update rule

Now in each step, the algorithm updates the membership matrix and the centers according to the following equations: $\forall\ 1 \leq j \leq k$ and $1 \leq i \leq n$

$$u_{ji}^{(t+1)} = \begin{cases} 1, & \text{if } X_i = Z_j^{(t)} \\ 0, & \text{if } X_i = Z_h^{(t)}, h \neq j \\ \dfrac{1}{\sum_{h=1}^{k} \left[\frac{d_{ji}(t)}{d_{hi}(t)}\right]^{\frac{1}{m-1}}} & \text{if } X_i \neq Z_j^{(t)} \text{ and } X_i \neq Z_h^{(t)}, 1 \leq h \leq k \end{cases} \tag{2}$$

$$z_{jl}^{(t+1)} = a_l^{(r)}, \tag{3a}$$

where

$$r = \arg \max_{1 \leq h \leq n_l} \sum_{\substack{i \\ x_{il} = a_l^{(h)}}}^{n} u_{ji}^{m(t+1)},$$

and

$$1 \leq h \leq n_l. \tag{3b}$$

Here as before each data point of X is described by p categorical attributes, $A_1, A_2, \ldots, A_p$. By n_l we denote the number of categories of attribute A_l, $1 \leq l \leq p$.

3. The weighted fuzzy k-modes (WF k-modes) algorithm

The performance of the fuzzy k-Modes algorithm is quite sensitive to initialization. Another main problem with the algorithm is that it treats all the variables equally while deciding upon the cluster memberships. However, the scenario is generally not the same for the real world categorical data. Often we can find the actual cluster structure hidden in some proper subspace of the entire feature set. Apart from increasing the computational overhead, emphasis on the relatively unimportant features may worsen the clustering performance, as their inclusion may prove detrimental to the actual cluster structure. We illustrate that presence of unimportant or noise variable do affect the clustering performance with an example of synthetic data. Here we consider a dataset of 50 data points, where the first variable is instrumental in demarcating the clusters, whereas the second variable is a noise

variable. We conduct 100 runs of the conventional fuzzy k-Modes clustering algorithm, on the single-dimensional (without any noise feature) and the two-dimensional (with noise feature) datasets. Then we use the Rand Index [18] to evaluate the performances of the clustering algorithms on the datasets. For a better visualization, we plot the Rand Index values after sorting them according to their values. Fig. 1 shows (in terms of Rand Index) that the introduction of noise indeed worsens the clustering performance.

In the proposed scheme, we introduce a weight vector in the conventional fuzzy k-Modes algorithm: $W=[w_1,w_2,\ldots,w_p]$, where $w_l=$ weight for the lth variable, $\forall l=1,2,\ldots,p$. Let β be the power of the attribute weight w_l. This β is a measure of emphasis on weights and as minimization of the objective function with respect to β is not possible (See Appendix for detailed explanation), it is fixed for $\forall l=1,2,\ldots,p$. The objective function to be minimized in case of the WFk-Modes algorithm is

$$O^{m,\beta}(U,Z,X,W)=\sum_{j=1}^{k}\sum_{i=1}^{n}u_{ji}^{m}d_{ji}^{W}, \tag{4}$$

where

$U=[u_{ji}]_{k\times n}$ is the membership matrix. $u_{ji}\in[0,1]$, $\forall j=1,2,\ldots,k$; $\forall i=1,2,\ldots,n$.

$X=(X_1,X_2,\ldots,X_n)$ i.e. the data matrix, where X_i is the ith data point(p dimensional)

$Z=(Z_1,Z_2,\ldots,Z_k)$ i.e. the cluster centers, where Z_j is centre of the jth cluster.

$d_{ji}^{W}=d^{W}(Z_j,X_i)=\sum_{l=1}^{p}\delta^{W}(z_{jl},x_{il})$ where z_{jl} is the lth term of Z_j and x_{il} is the l^{th} term of X_i.

$$\delta^{W}(z_{jl},x_{il})=\begin{cases}0, & z_{jl}=x_{il}\\ w_l^{\beta}, & z_{jl}\neq x_{il}\end{cases}$$

The minimization should be carried out subject to the following constraints:

$$\sum_{j=1}^{k}u_{ji}=1,\quad 1\le i\le n,$$

$$0<\sum_{i=1}^{n}u_{ji}<n,\quad 1\le j\le k \text{ and } \sum_{l=1}^{p}w_l=1$$

The steps of the WFk-Modes algorithm are presented below. Here $Z^{(t)}, U^{(t)}$, and $W^{(t)}$ respectively denote the centers, membership matrix and attribute weights at tth iteration.

(1) Randomly select k many data points $Z^{(1)}$ as modes. Randomly generate a set of initial weights $W^{(1)}$. Fix the fuzzifier value ($m=1.1$ in our case) and the Maximum number of Iteration ($T_{max}=50$ in our case). Determine membership matrix $U^{(1)}$ according to Eq. (2). Set iteration counter $t=1$.
(2) Determine $U^{(t+1)}$ according to Eq. (2). If $O^{m,\beta}\left(U^{(t+1)},Z^{(t)},X,W^{(t)}\right)=O^{m,\beta}\left(U^{(t)},Z^{(t)},X,W^{(t)}\right)$, stop. Otherwise go to step (c).
(3) Determine $Z^{(t+1)}$ according to Eq. (4). If $O^{m,\beta}\left(U^{(t+1)},Z^{(t+1)},X,W^{(t)}\right)=O^{m,\beta}\left(U^{(t+1)},Z^{(t)},X,W^{(t)}\right)$, stop. Otherwise go to step (d).
(4) Determine $W^{(t+1)}$ according to Eq. (5). Set $t=t+1$.If $O^{m,\beta}\left(U^{(t+1)},Z^{(t+1)},X,W^{(t+1)}\right)=O^{m,\beta}\left(U^{(t+1)},Z^{(t+1)},X,W^{(t)}\right)$or $t=T_{max}$, stop. Otherwise go to step (2).

3.1. Updating rule

At each step, update the membership matrix and the centers according to the following formulae:

for $\forall\ 1\le j\le k$ and $1\le i\le n$

$$u_{ji}^{(t+1)}=\begin{cases}1, & \text{if } X_i=Z_j^{(t)}\\ 0, & \text{if } X_i=Z_h^{(t)},h\neq j\\ \dfrac{1}{\sum_{h=1}^{k}\left[\frac{d_{ji}^{W(t)}}{d_{hi}^{W(t)}}\right]^{\frac{1}{m-1}}}, & \\ \text{if } X_i\neq Z_j^{(t)} \text{ and } X_i\neq Z_h^{(t)}, 1\le h\le k.\end{cases} \tag{5}$$

$$z_{jl}^{(t+1)}=a_l^{(r)},\ \text{where}\ \sum_{\substack{i\\ x_{il}=a_l^{(r)}}}^{n}u_{ji}^{m(t+1)}\ge\sum_{\substack{i\\ x_{il}=a_l^{(h)}}}^{n}u_{ji}^{m(t+1)},$$

$$1\le h\le n_l. \tag{6}$$

where each data point of X is described by p categorical attributes, $A_1,A_2,\ldots,A_p$.

$$w_l^{(t+1)}=\begin{cases}0 & \text{if } \Delta_l=0,\\ 1/\sum_{g=1}^{s}\left[\dfrac{\Delta_l^{(t+1)}}{\Delta_g^{(t+1)}}\right]^{\frac{1}{\beta-1}} & \text{if } \Delta_l\neq 0,\end{cases}$$

$$\Delta_l^{(t+1)}=\sum_{j=1}^{k}\sum_{i=1}^{n}u_{ji}^{m(t+1)}\delta\left(z_{jl}^{(t+1)},x_{il}\right), \tag{7}$$

where s is the number of variables where $\Delta_l\neq 0$.

The justification that the above mentioned updating rule indeed minimizes the objective function is provided in the form of the following three theorems:

Theorem 1. *Let $U=\hat{U}$ and $Z=\hat{Z}$ be fixed.*

(I) *When $\beta>1$ or $\beta\le 0$, $O^{m,\beta}(U,Z,X,W)$ is minimized iff*

$$\hat{w}_l=\begin{cases}0 & \text{if } \Delta_l=0\\ 1/\sum_{g=1}^{s}\left[\dfrac{\Delta_l}{\Delta_g}\right]^{\frac{1}{\beta-1}} & \text{if } \Delta_l\neq 0\end{cases}\qquad 1\le l\le p$$

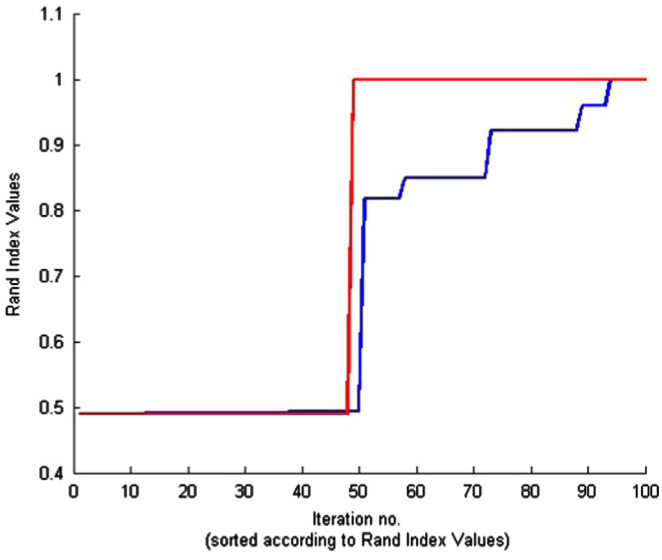

Fig. 1. Rand Index of 100 runs corresponding to the (a) main variable (without noise) (red) and (b) two-dimensional dataset, with the noise variable (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Where

$$\Delta_l = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^m \delta(z_{jl}, x_{il})$$

and s is the number of variables, where $\Delta_l \neq 0$.

(II) *When* $\beta = 1$, $O^{m,\beta}(U,Z,X,W)$ *is minimized iff*

$\hat{w}_{l'} = 1$ and $\hat{w}_l = 0, \quad l \neq l'$,

where $\Delta_{l'} \leq \Delta_l$ *for all l.*

Proof:

Part I.

$$\begin{aligned} O^{m,\beta}(U,Z,X,W) &= \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^m d_{ji}^W \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^m \sum_{l=1}^{p} \delta^W(z_{jl}, x_{il}) \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^m \sum_{l=1}^{p} w_l^\beta \delta(z_{jl}, x_{il}) \\ &= \sum_{l=1}^{p} w_l^\beta \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^m \delta(z_{jl}, x_{il}) = \sum_{l=1}^{p} w_l^\beta \Delta_l \end{aligned}$$

Here Δ_l's are p constants for fixed $\hat{U}$ and $\hat{Z}$. We want to minimize the objective function w.r.t to w_l's with the constraint $\sum_{l=1}^{p} w_l = 1$. If $\Delta_l = 0$, the $l-$th variable has a unique value in each cluster, which leads to a degenerate solution. Thus we are forced to assign $\hat{w}_l = 0$ to all variables, where $\Delta_l = 0$.

For the remaining $s(\leq p)$ variables, where Δ_l attains a non-zero value, we consider relaxing the minimization of the objective function via a Lagrange multiplier obtained by ignoring the constraint $\sum_{l=1}^{p} w_l = 1$. Let a be the multiplier and let the Lagrangian be

$$\Omega(W,a) = \sum_{l=1}^{s} w_l^\beta \Delta_l + a\left(\sum_{l=1}^{s} w_l - 1\right) \tag{8}$$

If $(\hat{W},\hat{a})$ minimizes $\Omega(W,a)$, the partial derivatives (gradients) of Ω, w.r.t. w_l's and a must vanish at $(\hat{W},\hat{a})$; thus

$$\left.\frac{\delta\Omega(W,a)}{\delta w_l}\right|_{(\hat{W},\hat{a})} = \beta \hat{w}_l^{\beta-1}\Delta_l + \hat{a} = 0, \quad 1 \leq l \leq p; \tag{9}$$

$$\left.\frac{\delta\Omega(W,a)}{\delta a}\right|_{(\hat{W},\hat{a})} = \sum_{l=1}^{s} \hat{w}_l - 1 = 0. \tag{10}$$

From (9), we obtain:

$$\hat{w}_l = \left(\frac{-\hat{a}}{\beta\Delta_l}\right)^{\frac{1}{\beta-1}} \tag{11}$$

Substitution of (11.) in. (10) gives,

$$\sum_{l=1}^{s}\left(\frac{-\hat{a}}{\beta\Delta_l}\right)^{\frac{1}{\beta-1}} = 1 \tag{12}$$

$$(-\hat{a})^{\frac{1}{\beta-1}} = 1\Big/\left[\sum_{g=1}^{s}\left(\frac{1}{\beta\Delta_g}\right)^{\frac{1}{\beta-1}}\right] \tag{13}$$

Another substitution of (13) in (11.) leads us to

$$\hat{w}_l = \left(\frac{1}{\beta\Delta_l}\right)^{\frac{1}{\beta-1}}\Big/\left[\sum_{g=1}^{s}\left(\frac{1}{\beta\Delta_g}\right)^{\frac{1}{\beta-1}}\right] = 1\Big/\left[\sum_{g=1}^{s}\left(\frac{\Delta_l}{\Delta_g}\right)^{\frac{1}{\beta-1}}\right]$$

Part II

For $\beta = 1$, the objective function becomes $\sum_{l=1}^{p} w_l\Delta_l$. It trivially follows that, $\sum_{l=1}^{p} w_l\Delta_l \geq \Delta_{l'}$ where $\Delta_{l'} \leq \Delta_l$ for all l, because,

$$\sum_{l=1}^{p} w_l(\Delta_l - \Delta_{l'}) \geq 0 \quad [\text{as all terms are non} - \text{negetive}]$$

$$\begin{aligned} \sum_{l=1}^{p} w_l(\Delta_l - \Delta_{l'}) &= \sum_{l=1}^{p} w_l\Delta_l - \sum_{l=1}^{p} w_l\Delta_{l'} \\ &= \sum_{l=1}^{p} w_l\Delta_l - \Delta_{l'}\sum_{l=1}^{p} w_l = \sum_{l=1}^{p} w_l\Delta_l - \Delta_{l'} \end{aligned}$$

Thus, setting $\hat{w}_{l'} = 1$ and $\hat{w}_l = 0, \forall l \neq l'$ will do the job for us.

For $\beta < 0$, attributes with higher importance will get lower weights, as here, $(1/w_l)^{|\beta|}$ is the effective weight corresponding to the lth attribute. This fact is reflected in the theoretical upgradation rule derived for general β. This particular range of β is also used earlier in literature by Huang et al. [14].□

Theorem 2. *Let* $W = \hat{W}$ *and* $Z = \hat{Z}$ *be fixed. When* $m > 1$, $O^{m,\beta}(U,Z,X,W)$ *is minimized if*

$$u_{ji} = \begin{cases} 1, & \text{if } X_i = Z_j \\ 0, & \text{if } X_i = Z_h,\ h \neq j \\ \dfrac{1}{\sum_{h=1}^{k}\left[\frac{d_{ji}^W}{d_{hi}^W}\right]^{\frac{1}{m-1}}}, & \text{if } X_i \neq Z_j \text{ and } X_i \neq Z_h, 1 \leq h \leq k \end{cases}$$

Proof:

$$O^{m,\beta}(U,Z,X,W) = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^m d_{ji}^W = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ji}^m d_{ji}^W$$

The minimization of the objective function can be performed by minimization of the function independently for each of the n data points. Hence a reduced problem could be the constrained (with same constraints) minimization of:

$$\sum_{j=1}^{k} u_{ji}^m d_{ji}^W \qquad \forall 1 \leq i \leq n$$

If $\exists j \in \{1,2,\ldots,k\}$ such that $X_i = Z_j$, then $0 = d_{ji}^W \geq d_{j'i}^W \quad \forall j' \neq j$; thus, setting $\hat{u}_{ji} = 1$ and $\hat{u}_{j'i} = 0 \quad \forall j' \neq j$will do the constrained minimization for us

Now assume $\nexists j \in \{1,2,\ldots,k\}$such that$X_i = Z_j$, then (for a fixed i) d_{ji}^W, $j \in \{1,2,\ldots,k\}$ are k constants for fixed $\hat{W}$ and $\hat{Z}$. Hence we are left with a similar situation as discussed in part 1of the proof of Theorem 1. Thus, we can have:

$$\hat{u}_{ji} = \frac{1}{\sum_{h=1}^{k}\left[\frac{d_{ji}^W}{d_{hi}^W}\right]^{\frac{1}{m-1}}} \qquad \forall 1 \leq j \leq k;\ \ \forall 1 \leq i \leq n.\square$$

Theorem 3. *Let* $U = \hat{U}$ *and* $W = \hat{W}$ *be fixed.* $O^{m,\beta}(U,Z,X,W)$ *is minimized iff*

$$z_{jl} = a_l^{(r)}, \text{ where, } \sum_{i,x_{il}=z_{jl}} u_{ji}^m \geq \sum_{i,x_{il}=a_l^{(h)}} u_{ji}^m, \qquad 1 \leq h \leq n_l$$

$$O^{m,\beta}(U,Z,X,W) = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^m d_{ji}^W$$

The objective function can be minimized by minimizing the function independently for each of the k cluster centers. Thus, the reduced problem $\forall 1 \leq j \leq k$ now becomes minimization of

$$\sum_{i=1}^{n} u_{ji}^m d_{ji}^W$$

$$\sum_{i=1}^{n} u_{ji}^m d_{ji}^W = \sum_{i=1}^{n} u_{ji}^m \sum_{l=1}^{p} \delta^W(Z_{jl}, X_{il})$$

$$= \sum_{i=1}^{n} u_{ji}^{m} \sum_{l=1}^{p} w_l^{\beta} \delta(z_{jl}, x_{il}) = \sum_{l=1}^{p} w_l^{\beta} \sum_{i=1}^{n} u_{ji}^{m} \delta(z_{jl}, x_{il})$$

$$= \sum_{l=1}^{p} w_l^{\beta} \left(\sum_{h=1}^{n_l} \sum_{i, x_{il} = a_l^{(h)}} u_{ji}^{m} - \sum_{i, x_{il} = z_{jl}} u_{ji}^{m} \right)$$

Clearly as far as the quantity within the bracket is concerned, the first term is a constant for fixed $\hat{U}$. The term outside the bracket is a constant for fixed $\hat{W}$. Hence, for minimization of the function, for each l, the second term, in the bracketed portion is to be maximized, i.e.

$$\sum_{i, x_{il} = z_{jl}} u_{ji}^{m} \geq \sum_{i, x_{il} = a_l^{(h)}} u_{ji}^{m}, \qquad 1 \leq h \leq n_l$$

Now, we discuss the relation between the respective minima of the conventional fuzzy k-modes and the proposed WFk-modes algorithm with the help of the following theorem:

Theorem 4. *Under the assumption that for a fixed dataset X, the minimization tasks are carried out on the feasible regions corresponding to each of the variables, the following two inequalities are observed:*

(1) $\forall \beta > 0$, $\min\ O^{m,\beta}(U,Z,X,W) < \min\ O^{m}(U,Z,X)$

(2) $\forall \beta < 0$, $\min\ O^{m,\beta}(U,Z,X,W) > \min\ O^{m}(U,Z,X)$

Proof: For a fixed dataset X, let us assume $\forall \beta > 0$. Now, $\left(\hat{U}, \hat{Z}\right)$ be any minima of $\min O^{m}(U,Z,X)$, i.e. $O^{m}\left(\hat{U}, \hat{Z}\right) = \min O^{m}(U,Z,X)$. Now, consider any weight vector $W = [w_1, w_2, \ldots, w_p]$, $w_l > 0$, $\forall l = 1, 2, \ldots, p$. Next we observe the following:

$$O^{m,\beta}\left(\hat{U}, \hat{Z}, X, W\right) = \sum_{j=1}^{k} \sum_{i=1}^{n} \hat{u}_{ji}^{m} \sum_{l=1}^{p} w_l^{\beta} \delta(\hat{z}_{jl}, x_{il})$$

$$< \sum_{j=1}^{k} \sum_{i=1}^{n} \hat{u}_{ji}^{m} \sum_{l=1}^{p} \delta(\hat{z}_{jl}, x_{il}) \left[\exists,\ i,j \text{ such that } \hat{Z}_j \neq X_i\right]$$

$$= O^{m}\left(\hat{U}, \hat{Z}, X\right) = \min\ O^{m}(U,Z,X)$$

Hence, from the above inequality, it follows that $\min O^{m,\beta}(U,Z,X,W) < \min O^{m}(U,Z,X)$. The proof of the theorem for $\beta < 0$ is similar.□

Next, we discuss the inter correlation of the distance metric and the weight vector with the help of the following Lemma.

Lemma 1. *For any fixed weight vector $W = [w_1, w_2, \ldots, w_p]$, $w_l > 0$, $\forall l = 1, 2, \ldots, p$, $d^{W} : C_{A_1,A_2,\ldots,A_p} \times C_{A_1,A_2,\ldots,A_p} \mapsto \mathbb{R}$ is a metric on $C_{A_1,A_2,\ldots,A_p}$ where*

$C_{A_1,A_2,\ldots,A_p} = \{X = [x_1, x_2, \ldots, x_p] \mid x_l \in DOM(A_l) \forall\ l = 1, 2, \ldots, p\}$

Proof:

Let $S_1, S_2, S_3 \in C_{A_1,A_2,\ldots,A_p}$. Then

(1) $d^{W}(S_1,S_2) = \sum_{l=1}^{p} \delta^{W}(s_{1l}, s_{2l}) \geq 0$, equality holds iff, $s_{1l} = s_{2l}$, $\forall l = 1, 2, \ldots, p$; *i.e.* $S_1 = S_2$.

(2) $d^{W}(S_1,S_2) = \sum_{l=1}^{p} \delta^{W}(s_{1l}, s_{2l}) = \sum_{l=1}^{p} \delta^{W}(s_{2l}, s_{1l}) = d^{W}(S_2,S_1)$.

(3) $d^{W}(S_1,S_2) + d^{W}(S_2,S_3) = \sum_{l=1}^{p} \left(\delta^{W}(s_{1l}, s_{2l}) + \delta^{W}(s_{2l}, s_{3l})\right) \geq \sum_{l=1}^{p} \delta^{W}(s_{1l}, s_{3l}) = d^{W}(S_1,S_3)$.

These three properties prove the claim of the Lemma.

In case of $w_l = 0$ for some $l \in \{1, 2, \ldots, p\}$, the contribution of the corresponding attributes in the objective function is identically equal to 0, so we can just neglect the corresponding attributes and concentrate on the attributes with $w_l > 0$. Now, from Lemma 1 it is evident that the weight vector induces a metric on the space of the data points under consideration and the proposed algorithm is a fuzzy k-modes algorithm with this new distance metric with an automated upgradation policy for the weight vector inducing the distance metric. Being an integrated part of the distance metric under consideration, the weight vector is not a part of membership vectors.

4. Experimental results and performance measures

In Table 1, we provide a brief description of the validity indices and the accuracy measures that we use for comparison of the clustering performances of the conventional fuzzy k-Modes algorithm [2] and our newly proposed algorithm. Here we use Rand Index [18] for measuring the degree of matching between the original partition and the obtained hard partition from the clustering algorithms. A hard partition is obtained by assigning each data point to the cluster, in which it has highest membership. In case of a draw, we choose the last cluster. For fuzziness measurements, Partition Coefficient [19] and Partition Entropy [20] are used. (Table 2)

The notations, used in the functional form of Rand Index are defined as follows:

The original crisp partition of the dataset under consideration is denoted by $\mathbb{P}_{ORI}$. The obtained crisp partition is denoted by $\mathbb{P}_{RES}$. For calculation purpose, every pair of points of dataset is tagged as follows:

(1) *SS* if both of the points belong to the same cluster in $\mathbb{P}_{ORI}$ and $\mathbb{P}_{RES}$.
(2) *DD* if both of the points belong to two different clusters in $\mathbb{P}_{ORI}$ and $\mathbb{P}_{RES}$.
(3) *SD* if the two points belong to the same cluster in $\mathbb{P}_{ORI}$ and to different clusters in $\mathbb{P}_{RES}$.
(4) *DS* if the two points belong to different clusters in $\mathbb{P}_{ORI}$ and to the same cluster in $\mathbb{P}_{res}$.

Let c_1, c_2, c_3 and $c_4$be the number of points tagged as *SS*, *DD*, *SD*, and *DS* respectively. To test the clustering performance of the two algorithms, we choose one synthetic and two real world data sets from the UCI repository [17]. For each dataset we consider 100 runs of the weighted fuzzy k-Modes clustering algorithm, with the coefficient of weights i.e. β varying from -10 to 10, excluding 1 (As by Theorem 3, for $\beta = 1$, all weights will be assigned to the variable having least variability and others will get 0 weight, which is an undesirable situation). We observe that the conventional fuzzy k-Modes algorithm can be seen as a special case of its weighted version with coefficient of weight i.e. $\beta = 0$. We compute Rand Index, Partition Coefficient, and Partition Entropy of the clustering performance of the new algorithm. We then carry out a statistical analysis of the data to show that our proposed

Table 1
Mathematical description of the performance measures.

Performance measure	Functional description
Partition Coefficient (PC)	$V_{pc}(U) = \left(\sum_{j=1}^{k} \sum_{i=1}^{n} u_{ji}^{2} \right) / n$
Partition Entropy (PE)	$V_{pe}(U) = \left[\sum_{j=1}^{k} \sum_{i=1}^{n} (u_{ji} \log\ u_{ji}) \right] / n$
Rand Index (Rand)	$\frac{(c_1 + c_4)}{(c_1 + c_2 + c_3 + c_4)} = \frac{(c_1 + c_4)}{\frac{n(n-1)}{2}}$

Table 2
Functional description of the performance measures.

Performance indicator	Measured property	Optimal partition
Partition Coefficient (PC)	Fuzziness of the partition	$Max(V_{pc})$[19-22]
Partition Entropy (PE)	Fuzziness of the partition	$Min(V_{pe})$[19-22]
Rand Index (Rand)	Matching between the actual crisp partition and the obtained crisp partition	*Max*(Rand Index) [18]

Table 3
Robustness of algorithm proposed in [24] with respect to initial weights.

Cluster specific weighting		
Initial weights	**Final weights**	**Rand Index**
$\begin{bmatrix}0.94566 & 0.0543\\ 0.612 & 0.388\end{bmatrix}$	$\begin{bmatrix}0.8889 & 0.1111\\ 0.7917 & 0.2083\end{bmatrix}$	0.754
$\begin{bmatrix}0.94566 & 0.0543\\ 0.374 & 0.6258\end{bmatrix}$	$\begin{bmatrix}0.8889 & 0.1111\\ 0.7917 & 0.2083\end{bmatrix}$	0.738
$\begin{bmatrix}0.94566 & 0.0543\\ 0.034 & 0.966\end{bmatrix}$	$\begin{bmatrix}0.8947 & 0.1053\\ 0.8182 & 0.1818\end{bmatrix}$	0.73
$\begin{bmatrix}0.612 & 0.388\\ 0.374 & 0.6258\end{bmatrix}$	$\begin{bmatrix}0.8947 & 0.1053\\ 0.8182 & 0.1818\end{bmatrix}$	0.711
$\begin{bmatrix}0.612 & 0.388\\ 0.034 & 0.966\end{bmatrix}$	$\begin{bmatrix}0.8947 & 0.1053\\ 0.8182 & 0.1818\end{bmatrix}$	0.71
$\begin{bmatrix}0.374 & 0.6258\\ 0.034 & 0.966\end{bmatrix}$	$\begin{bmatrix}0.8947 & 0.1053\\ 0.8182 & 0.1818\end{bmatrix}$	0.694

Table 4
Robustness of the proposed algorithm with respect to initial weights.

WF*k*-modes		
Initial weights	**Final weights**	**Rand Index**
(0.94566, 0.0543)	(0.648, 0.351)	1
(0.612, 0.388)	(0.648, 0.351)	1
(0.374, 0.62579)	(0.648, 0.351)	0.984
(0.034, 0.966)	(0.648, 0.351)	0.981

algorithm outperformed the fuzzy *k*-Modes algorithm in terms of fuzziness of the partition and showed non decreasing performance, in terms of the Rand Index values. We plot the histograms of the Rand Index, Partition Coefficient, and Partition Entropy values which provide evidences in the favor of the better performance of the newly proposed algorithm (histograms are drawn with "value of the performance indicator" and "β-value" along *X*-axis respectively). Finally we carry out paired Wilcoxon's rank sum test [23] with respect to $\beta=0$ and other non-zero values of β, to check for statistically significant improvement of performance. In Fig. A2 of the Appendix we show the VAT pictures of three of the used datasets (synthetic, soybean, and zoo) for guaranteeing the presence of actual cluster structure in the data.

Example 1. DATASET: Synthetic data
ALGORITHM: WF*k*-Modes and fuzzy *k*-Modes.

This dataset contains 2 variables and 50 points, divided into 2 clusters, each cluster containing 25 points each. The first variable is the main variable, which determines the cluster structure, while the randomly generated second variable acts as a noise variable. The first variable of the first dataset is a binary variable taking the 2 values: 1and 2. The second variable is a random natural number between 1 and 10. Among the 50 data points, 25 data points with first variable value 1, belong the first cluster, and the rest of the 25 data points with first variable value 2 belong to the second cluster. This dataset can be interpreted as a categorical data with 2 categorical attributes, A_1 and A_2. Here, $DOM(A_1)=\{1,2\}$ and $DOM(A_2)=\{1,2,\ldots,10\}$. Any categorical data can be translated in this format. This data set is equivalent to the categorical data with two attributes, A_1 and A_2, where, $DOM(A_1)=\{"red","blue"\}$ and $DOM(A_2)=\{"A","B",\ldots,"J"\}$. As expected the first variable plays instrumental role in detection of the appropriate clustering structure. The second variable is the noise variable, and if given equal weight as the first variable during clustering, the clustering performance will be worsened. In Fig. A1, provided in the Appendix, we show the frequencies of the values of the two variables; the noisy nature of the second variable becomes prominent from the plot. The main clustering structure is hard to recover if the two variables are given same weights. We now demonstrate that the WF*k*-Modes algorithm was able to detect the noise variable and recover the clustering pattern successfully.

In Fig. 2(a)–(c) we provide the average values of the Rand Index, Partition Coefficient, and Partition Entropy along with the table, over the 30 independent runs with randomly initialized modes and weights (for each run, initialization was fixed for different values of β). We see that the average value of Rand Index for the conventional fuzzy *k*-Modes algorithm is lower than that of its weighted counterpart for all β's. Our newly proposed algorithm also outperforms the conventional fuzzy *k*-modes in terms of mean V_{PC} and V_{PE} too.

From the histogram plots (Fig. 6(a) and (b)) it is evident that, the conventional fuzzy *k*-Modes never provides a perfect clustering of the data, where as our algorithm (for all considered nonzero values of β) fails to give the perfect clustering only in a single case in all of the 30 runs. As far as the fuzziness of the partition is concerned, the picture is pretty much same there also. The conventional fuzzy *k*-Modes has V_{PC} mostly in $0.9-0.925$ and one in $0.625-0.65$ (Fig. 6(c)), but our algorithm's V_{PC} for all the concerned values of $\beta\neq 0$, are above 0.95 (Fig. 6(d)) region and one observation around 0.7 (for some values of β). For V_{PE} too, we have the same pattern, they are mostly below 0.05 (Fig. 6(f)) region for WF*k*-modes and in $0.1-0.15$ (Fig. 6(e)) region for conventional *k*-modes.

Even for the pathological case, our algorithm performed reasonably well for some βs. As, all the histograms for the considered non zero values of β, look similar, we just provided the histogram where the effect of bad initialization is observed; The cases with *all* perfect clustering gives the histogram as just a single bar at 1 [for Rand Index], $(0.99, 1)$ [for V_{pc}] and $(0, 0.2)$ [for V_{pc}].

For further investigation into the data, we undertook the Wilcoxon's Rank Sum test (paired) (also known as the Mann–Whitney *U* test), to test if our algorithm provides us with statistically significant median results. We report the *P*-values in Table 5. In the tables corresponding to *P*-values the column headings represent for which indices we are carrying out the rank-sum test (Rand=Rand Index, PC=Partition Coefficient, PE=Partition Entropy) and the alternative hypothesis for that

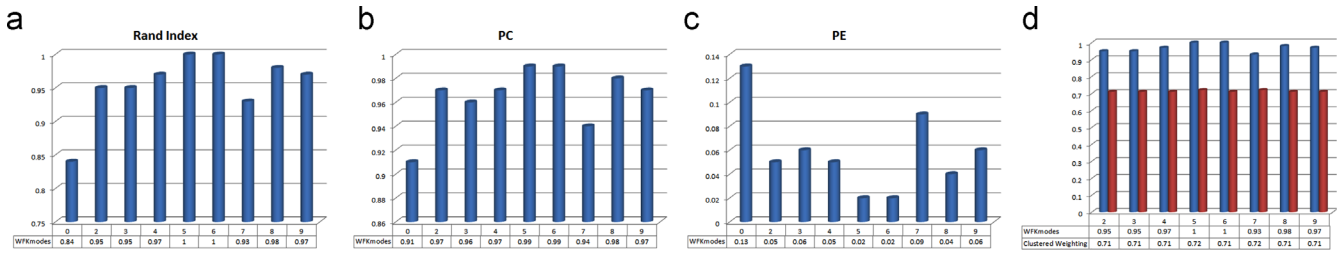

Fig. 2. For Synthetic Dataset, for different values of βmean values of (a) Rand Index, (b) V_{pc}, (c)V_{pe}; (d) Mean Rand Index for WF*k*-modes and Cluster Specific Weighting.

Table 5
P-Values for synthetic dataset for all the performance measures.

β	Rand (>0)	PC (>0)	PE (<0)
2	0.0009	0.001	0.001
3	0.0009	0.001	0.001
4	0.0001	0.0002	0.0002
5	$4E-07$	$9E-07$	$9E-07$
6	$4E-07$	$9E-07$	$8E-07$
7	0.005	0.006	0.006
8	$9E-06$	$1E-05$	$1E-05$
9	$1E-04$	$2E-04$	$1E-04$

Table 6
Robustness (in terms of Rand Index) of WFk-modes with increment in number of noise variables.

	Dimension of the data					
β	5	10	25	50	75	100
0	0.74	0.67	0.58	0.52	0.57	0.53
2	1	0.89	0.95	0.99	0.99	1
3	0.92	1	0.95	0.99	0.99	0.96
4	0.9	1	1	1	1	0.94
5	1	1	1	1	1	0.77
6	0.83	0.92	1	0.99	0.92	0.72
7	0.9	0.91	0.96	0.99	0.81	0.74
8	0.9	1	0.96	0.99	0.83	0.72
9	0.92	1	0.99	0.87	0.86	0.64

Table 7
Robustness (in terms of V_{PC}) of WFk-modes with increment in number of noise variables.

	Dimension of the data					
β	5	10	25	50	75	100
0	0.85	0.82	0.73	0.7	0.69	0.66
2	0.99	0.9	0.94	0.94	0.92	0.85
3	0.95	1	0.96	0.912	0.89	0.85
4	0.96	1	1	0.97	0.91	0.88
5	1	1	1	0.9	0.97	0.77
6	0.92	0.94	1	0.94	0.92	0.73
7	0.96	0.94	0.93	0.95	0.83	0.75
8	0.96	1	0.93	0.93	0.76	0.72
9	0.96	1	0.99	0.84	0.82	0.7

tests ((>0)=alternative hypothesis: true location shift from 0 for corresponding β is greater than 0, (<0)=alternative hypothesis: true location shift from 0 for corresponding β is less than 0). From Table 5, we can find out that indeed all the values of β provide us with statistically significant amount of improvement in the Rand Index, $V_{PC}$and V_{PE}.

In [24] an automated feature weighting was introduced for the FCM algorithm, if we just adopt the idea, and replace the fuzzy c-means part by the *k*-modes part, we DO NOT end up with our proposed algorithm. In that case, feature weight was provided for each of the variables, in each cluster, so single variable may have different weights in different cluster, which is not going to happen in our case. We adjudge the feature and provide its weight, according to its importance in recovering the cluster structure and then cluster the data according to that, but the weighting scheme in the aforementioned paper, is directly linked with the obtained partition, i.e. importance of any variable, within a particular cluster, but not in whole clustering process. As our algorithm speaks of a unique weight for each of the variables under consideration, it also gives rise to the idea of dimensionality reduction but that is not guaranteed in the case with aforementioned algorithm. We also provide a comparison of results to support the claim of better performance of our algorithm. Fig. 2(d) clearly shows that our algorithm outperforms the afore mentioned algorithm, for all considered $\beta \neq 0$.We go further for comparing the robustness of our algorithm with the concerned algorithm w.r.t initialization.

Table 8
Robustness (in terms of V_{PE}) of Wfk-modes with increment in number of noise variables.

	Dimension of the data					
β	5	10	25	50	75	100
0	0.24	0.28	0.41	0.46	0.48	0.51
2	0	0.16	0.10	0.096	0.13	0.245
3	0.08	0	0.03	0.143	0.194	0.244
4	0.06	0	0	0.056	0.145	0.205
5	1	0	0	0.139	0.076	0.36
6	0.13	0.09	0	0.145	0.144	0.42
7	0.07	0.1	0.05	0.097	0.274	0.387
8	0.07	0	0.06	0.137	0.409	0.44
9	0.06	0	0.01	0.312	0.291	0.47

For testing the robustness w.r.t initiation weights, we now vary the choice of weights, i.e. the ratio of weight for the first and the second variable. We do this for 4 sets of initial weights for our algorithm, and for all the possible 6 pairs of weights possible for the algorithm proposed in [24]. Tables 3–4 show that our algorithm outperforms its competitor by a significant margin.

We now want to test the robustness of the proposed algorithm towards increment of noise in the data and compare with that of the conventional *k*-Modes algorithm.

For this purpose, we gradually increase the number of noise variables and we try to find to what extent our algorithm is able to find the proper cluster structure under these setups. We tabulate the mean Rand Index, V_{PC} and V_{PE} over 30 runs for the cases where the dimension of the dataset is 5, 10, 25, 50, 75, 100 (in each case, all except the first variable was noise variable). From Tables 6–8, it becomes evident that even in presence of very high degree of noise for some choices of

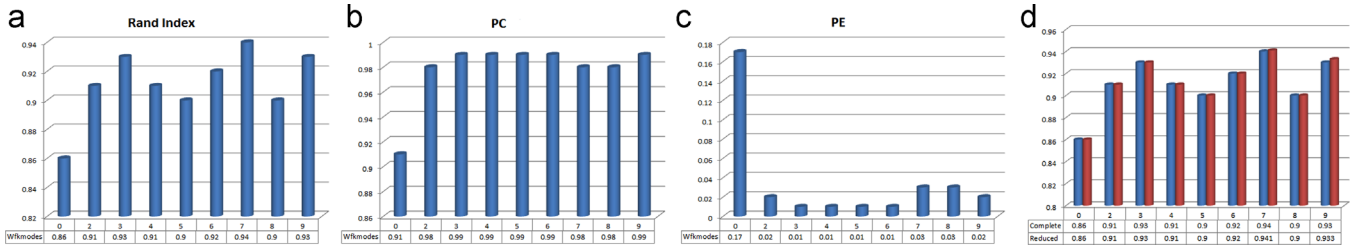

Fig. 3. For Soybean Dataset, for different values of β, mean values of (a) Rand Index, (b) V_{pc}, (c)V_{pe}, and (d) mean Rand Index for WF*k*-modes in complete and reduced data.

Table 9
P-values for soybean dataset for all the performance measures.

β	Rand(>0)	PC(>0)	PE(<0)
2	0.004	$2E-06$	$8E-07$
3	0.0008	$9E-10$	$9E-10$
4	0.02	$9E-10$	$9E-10$
5	0.03	$9E-10$	$9E-10$
6	0.005	$9E-10$	$9E-10$
7	0.002	$2E-06$	$7E-07$
8	0.01	$1E-05$	$5E-06$
9	0.0002	$9E-10$	$9E-10$

β (depending on the initialization and the number of noise variables) we are getting perfect clustering for all of 30 runs, where as the conventional *k*-modes performs the worst consistently in all the cases.

For fulfillment of the main purpose of performance comparison of the newly proposed algorithm with conventional fuzzy *k*-modes algorithm, the datasets (clustering problems) were chosen such that the ground truth (class) for each of them were known apriori. The datasets with varied number of attributes, instances (data points) and actual classes were chosen for sake of fair comparison.

Example 2. DATASET: Soybean Data
ALGORITHM: WF*k* -Modes and fuzzy *k*-Modes.

Soybean data: The data set contains 47 instances, 35 attributes, and 4 clusters. Out of 35 we shall use selected 21 attributes, because, the other 14 have only one category. Fig. 3 (a)–(c) shows that, mean Rand Index and mean V_{PC} was increased to a significant extent for all nonzero βs under consideration. Mean V_{PE} is decreased for all β under consideration. So, all non-zero βs considered here, show us an all-round improvement in mean performance.

From the histograms (Fig. 6(g) and (h)) we clearly see that the Rand Index values obtained in *k*-modes algorithm, are more around $0.8-0.85$ and relatively small number of observations above 0.95.

On the other hand, most of the observations in case of WF*k*-modes is beyond 0.98. We observe the same pattern in histograms of V_{PC} (Fig. 6(i) and (j)) and V_{PE} (Fig. 6(k) and (l)) too. The *P*-values (Table 9) also provide proof in favor of statistically significant amount of overall improvement in performance of our algorithm for $\beta=8$ and 9.

Example 3. DATASET: Zoo data
ALGORITHM: WF*k* -Modes and fuzzy *k*-Modes.

Zoo data: The data set contains 101 instances, 17 attributes, and 7 clusters. The first attribute contains the name of the animals (insignificant) hence omitted during clustering. Here 15 attributes are Boolean in nature and one attribute takes values in $\{0,2,4,5,6,8\}$. We proceed as earlier and tabulate the mean values of performance measures. Fig. 4(a)–(c)shows that, mean Rand Index was increased to a significant extent for $\beta=2,\ 6-9$, mean V_{PC} is increased for $\beta=9$, mean V_{PE} is decreased for $\beta=2,\ 6-9$. So, $\beta=8,\ 9$ show us an all-round improvement in mean performance.

From the histograms (Fig. 7.(a)–(d)) we see that the Rand Index values obtained in *k*-modes algorithm, are more around 0.85 and very little observations above 0.9 and no observation beyond 0.95. On the other hand, most of the observations in case of WF*k*-modes is above 0.95. Though *k*-Modes algorithm depends much on the initialization, we can say that our algorithm is going to give better average performance than conventional *k*-modes in general. The *P*-values (Table 10) also provide proof in favor of better performance of our algorithm. Improvements in V_{PC} (Fig. 7(c) and (d)) and V_{PE} (Fig. 7(e) and (f)) were also observed.

Example 4. DATASET: Mushroom Data
ALGORITHM: WF*k* -Modes and fuzzy *k*-Modes.

Mushroom data: The data set contains 8124 instances, 23 attributes, and 2 clusters. The first attribute contains the nature of the mushrooms (the actual classes) hence omitted during clustering. The number of missing attribute value us 2480, so we omit the corresponding data points. We proceed as earlier and tabulate the mean values of performance measures. Fig. 5 (a)–(c) shows that, mean Rand Index was increased to a significant extent for $\beta=3,\ 4,9$ mean V_{PC} is increased for $\beta=2-7,9$, mean V_{PE} is decreased for $\beta=2-7,\ 9$. So, $\beta=3,4,\ 9$ show us an all-round improvement in mean performance.

From the histograms (Fig. 7(g) and (h)) we see that the Rand Index values obtained in *k*-modes algorithm, are around 0.73 and 0.5. On the other hand, all of the observations in case of WF*k*-modes (for some selected β) arearound 0.74.As far asV_{PC} and V_{PE} are concerned WF*k*-modes provided with values around 0.95 and 0.05 (Fig. 7(j) and (l)) respectively, whereas conventional *k*-modes achieved results as good as that less frequently (Fig. 7(i) and (k)). The *P*-values (Table11) also provide proof in favor of significantly improved performance of our algorithm. All round and statistically significant improved performance was observed in case of $\beta=3,4,6,9$.

4.1. Discussion on dimensionality reduction

Our algorithm associated a weight to each of the variables under consideration, according to their relative importance in recovering the cluster structure from the data. This provides us with a way of reducing the dimensionality of the data with little or no loss of cluster structure. If in a clustering algorithm, we get zero weight (in the final weight distribution after the convergence of the algorithm) corresponding to some variables, we can

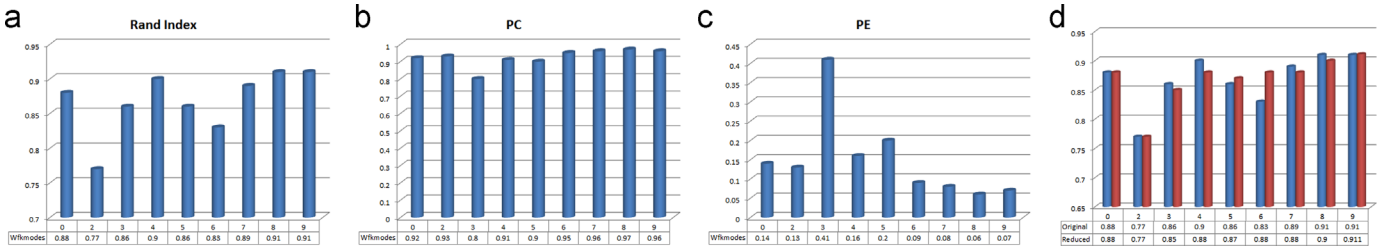

Fig. 4. For Zoo Dataset, for different values of β, mean values of (a) Rand Index, (b) V_{pc}, (c) V_{pe}, and (d) mean Rand Index for WF*k*-modes in complete and reduced data.

Table 10
P-Values for Zoo dataset for all the performance measures.

β	Rand(>0)	PC(>0)	PE(<0)
2	0.99	0.14	0.19
3	0.86	0.99	0.99
4	0.059	0.48	0.48
5	0.46	0.55	0.58
6	0.96	0.002	0.002
7	0.22	0.02	0.0009
8	0.038	$5E-06$	$3E-06$
9	0.045	$7E-05$	$6E-05$

hope that these variables do not have much roll in recovering the cluster structure, so the removal of these noise variables from the dataset may enhance the clustering performance and also reduce the computational complexity of the algorithm. To show this is indeed the case, for each of the datasets from the β's, which produced the best mean Rand Index in 30 runs, we randomly choose one of the better clustering results, and consider its associated weights. We remove the variables, corresponding to zero weight and perform the clustering again. We perform this for each one of the real world datasets—Soybean, Zoo and Mushroom datasets. The case specific results and discussion is provided below:

Soybean Data: In this case, we achieved best mean Rand Index value for $\beta=9$, so according to the scheme mentioned above, we removed variable number 2, 26 and 27 from the main dataset and performed clustering with both conventional *k*-modes and newly proposed algorithm. We found out (Fig. 3(d)) that average Rand Index value was at par with that corresponding to the main data set. Even slight improvement in mean Rand Index value was observed in some cases.

Zoo Data: In this case, we achieved best mean Rand Index value for $\beta=9$, so according to the scheme mentioned above, we removed variable number9 and 10 from the main dataset and performed clustering with both conventional *k*-modes and newly proposed algorithm. We found out (Fig. 4(d)) that the average Rand Index value was more or less same as that corresponding to the earlier dataset. Even slight improvement in best mean Rand Index value was observed.

Mushroom Data: In this case, we achieved best mean Rand Index value for $\beta=3$, so according to the scheme mentioned above, we removed variable number 17 from the main dataset and performed clustering with both conventional *k*-modes and WF*k*-modes algorithm. We found out (Fig. 5(d)) that average Rand Index value remained almost same as that corresponding to the main data set. Even slight improvement in best mean Rand Index value was observed.

Thus it can be said that, if we remove the less important variables, not much information is lost, sometimes, finding the right cluster structure becomes easier.

4.2. Discussion on fundamental difference between the optimization tasks carried out in case of conventional Fuzzy k-modes and its weighted counterpart

In this subsection, for sake of notational simplicity, we define the following subsets:

$$\mathscr{M}_{kn}=\left\{U=[u_{ji}]\text{is any real } k\times n \text{ matrix}\middle|\begin{array}{r}\sum_{j=1}^{k}u_{ji}=1,\ 1\le i\le n,\\ 0<\sum_{i=1}^{n}u_{ji}<n,\ 1\le j\le k\end{array}\right\}$$

$\mathscr{H}^p$ is the standard $(p-1)$-simplex (or unit $(p-1)$-simplex), i.e. it is the subset of $\mathbb{R}^p$, given by,

$$\mathscr{H}^p=\left\{(w_1,w_2,\ldots,w_p)\in\mathbb{R}^p\middle|\sum_{l=1}^{p}w_l=1, w_l\ge 0,\ l=1,2,\ldots,p\right\}$$

For a fixed set of given dataset X,

$$O^{m,\beta}:\ \mathscr{M}_{kn}\times C^{k}_{A_1,A_2,\ldots,A_p}\times\mathscr{H}^p\mapsto\mathbb{R}$$

$$\forall U\in\mathscr{M}_{kn},\ Z=(Z_1,Z_2,\ldots,Z_k),\ Z_j\in C_{A_1,A_2,\ldots,A_p},\ \ \forall j=1,2,\ldots k,;W\in\mathscr{H}^p$$

$$O^{m,\beta}(U,Z,W)=O^{m,\beta}(U,Z,X,W)=\sum_{j=1}^{k}\sum_{i=1}^{n}u_{ji}^{m}\sum_{l=1}^{p}w_l^{\beta}\delta(z_{jl},x_{il});$$

On the other hand,

$$O^{m}:\ \mathscr{M}_{kn}\times C^{k}_{A_1,A_2,\ldots,A_p}\mapsto\mathbb{R}$$

$$\forall U\in\mathscr{M}_{kn},\ Z=(Z_1,Z_2,\ldots,Z_k),Z_j\in C_{A_1,A_2,\ldots,A_p},\ \ \forall j=1,2,\ldots k$$

$$O^{m}(U,Z)=O^{m}(U,Z,X)=\sum_{j=1}^{k}\sum_{i=1}^{n}u_{ji}^{m}\sum_{l=1}^{p}\delta(z_{jl},x_{il})$$

Hence, the objective functions in case of conventional *k*-modes and the WF*k*-modes are two completely different functions, defined on two separate domains. The optima of these two objective functions are inherently different, which in turn guarantees different clustering performances.

4.3. Discussion on fuzziness of the obtained clustering with WFk-modes

The theoretical study of the fuzziness of the obtained clustering at the optimum of the objective function of WF*k*-modes and its comparison with that corresponding to its unweighted counterpart is out of the scope of this present article. Here, from the results in the experimental section, we observe that fuzziness generally decreases for WF*k*-modes in case of synthetic (Fig. 2(b) and (c)) and soybean dataset (Fig. 3(b) and (c)); whereas for zoo dataset (Fig. 4(b) and (c)) and mushroom (Fig. 5(b) and (c)) dataset, we get a mixed trend going by the average value of the cluster validity indices(V_{PC}, V_{PE}).

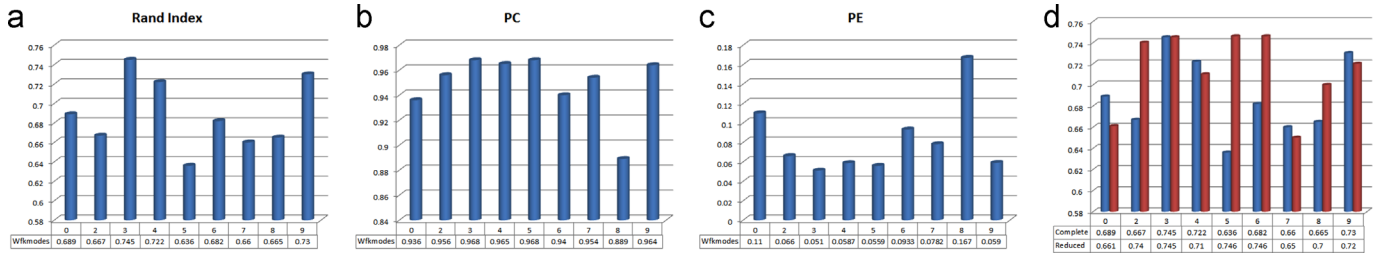

Fig. 5. For Mushroom Dataset, for different values of β, mean values of (a) Rand Index, (b) V_{pc}, (c)V_{pe}, and (d) mean Rand Index for WF*k*-modes in complete and reduced data.

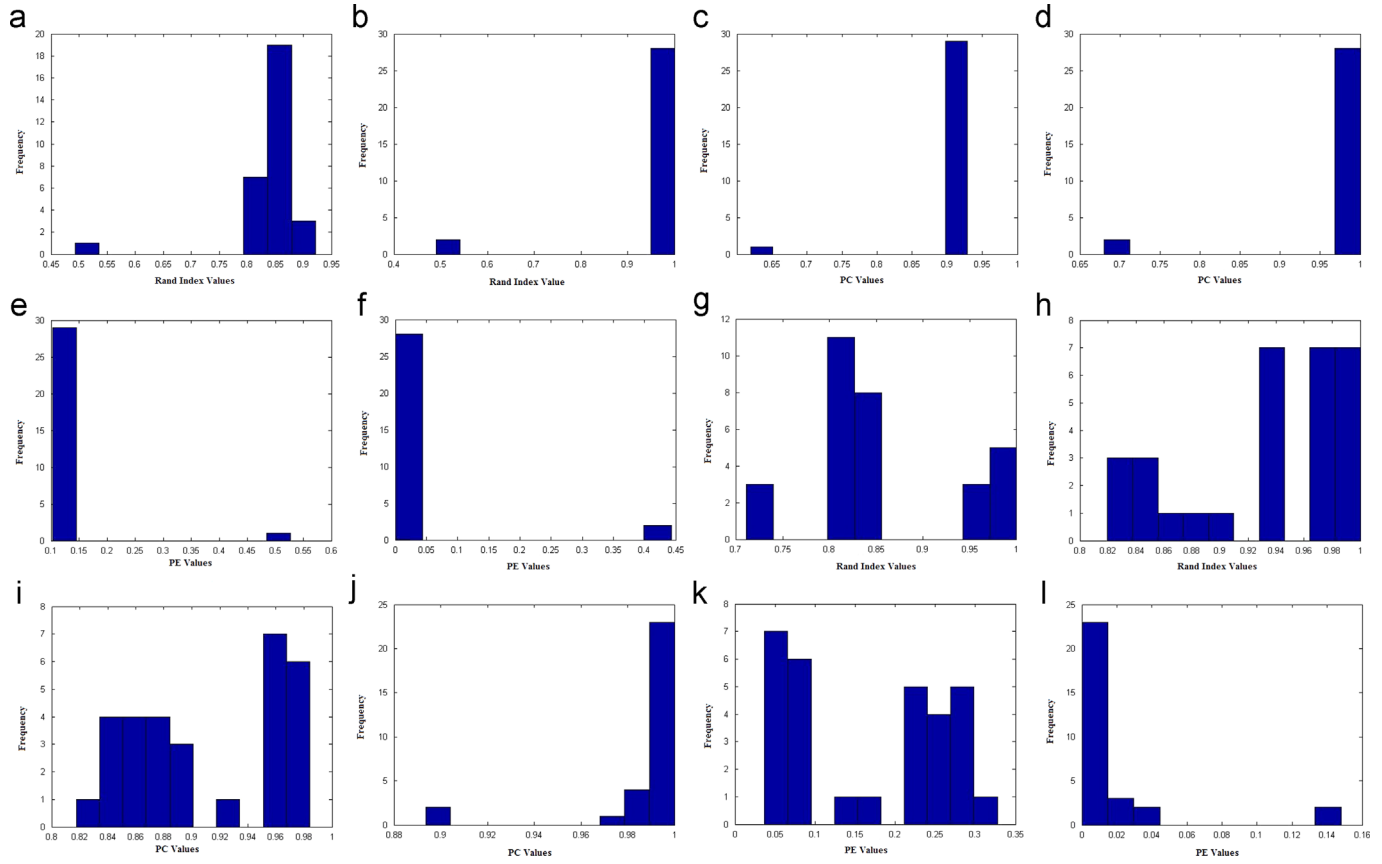

Fig. 6. For synthetic dataset, Rand Index corresponding to (a) $\beta = 0$ (b) $\beta = 9$; V_{pc}corresponding to (c) $\beta = 0$ (d) $\beta = 9$;V_{pe}corresponding to (e) $\beta = 0$, (f) $\beta = 9$. for Soybean Dataset, Rand Index corresponding to (g) $\beta = 0$ (h) $\beta = 9$; V_{pc}corresponding to (i) $\beta = 0$ (j) $\beta = 9$; V_{pe}corresponding to (k) $\beta = 0$ (l) $\beta = 9$.

Table 11
P-Values for Mushroom dataset for all the performance measures.

β	**Rand(> 0)**	**PC(> 0)**	**PE(< 0)**
2	0.69	0.002	0.002
3	$2.1E-06$	1.2E−06	$8.7E-07$
4	0.0007	0.0005	0.0004
5	0.61	9.7E−05	0.0003
6	0.01	0.005	0.005
7	0.31	0.11	0.11
8	0.34	0.67	0.65
9	$8.2E-05$	0.0001	0.0001

4.4. Modification of algorithm for unknown number of clusters

In this article the weighted Fuzzy *k*-modes algorithm is developed for clustering problems, where the number of "actual" or "true" classes is known, but in real world scenario finding the actual number of clusters maybe a difficult problem even with VAT pictures. With help of a proper cluster validity measure, our newly proposed algorithm can be modified to be used to in this new scenario also. In this case, we first fix the maximum ($M_{max} \leq n$) and the minimum ($M_{min} \geq 2$) number of cluster possible. Next, for $\forall k$, such that $M_{min} \leq k \leq M_{max}$, we perform the usual WF*k*-modes clustering with number of clusters fixed at k and note the value of the chosen cluster validity index after convergence for each k. Finally, we choose the k and the corresponding clustering result, for which the best cluster validity index value was achieved. Though the best choice of this particular cluster validity index and its relationship with the weight vector is beyond the scope of this article and can be taken up as a promising future avenue of research.

4.4.1. Discussion on the computational complexity

Total number of data points $= n$
Total number of clusters $= k$
Total number of attributes $= p$

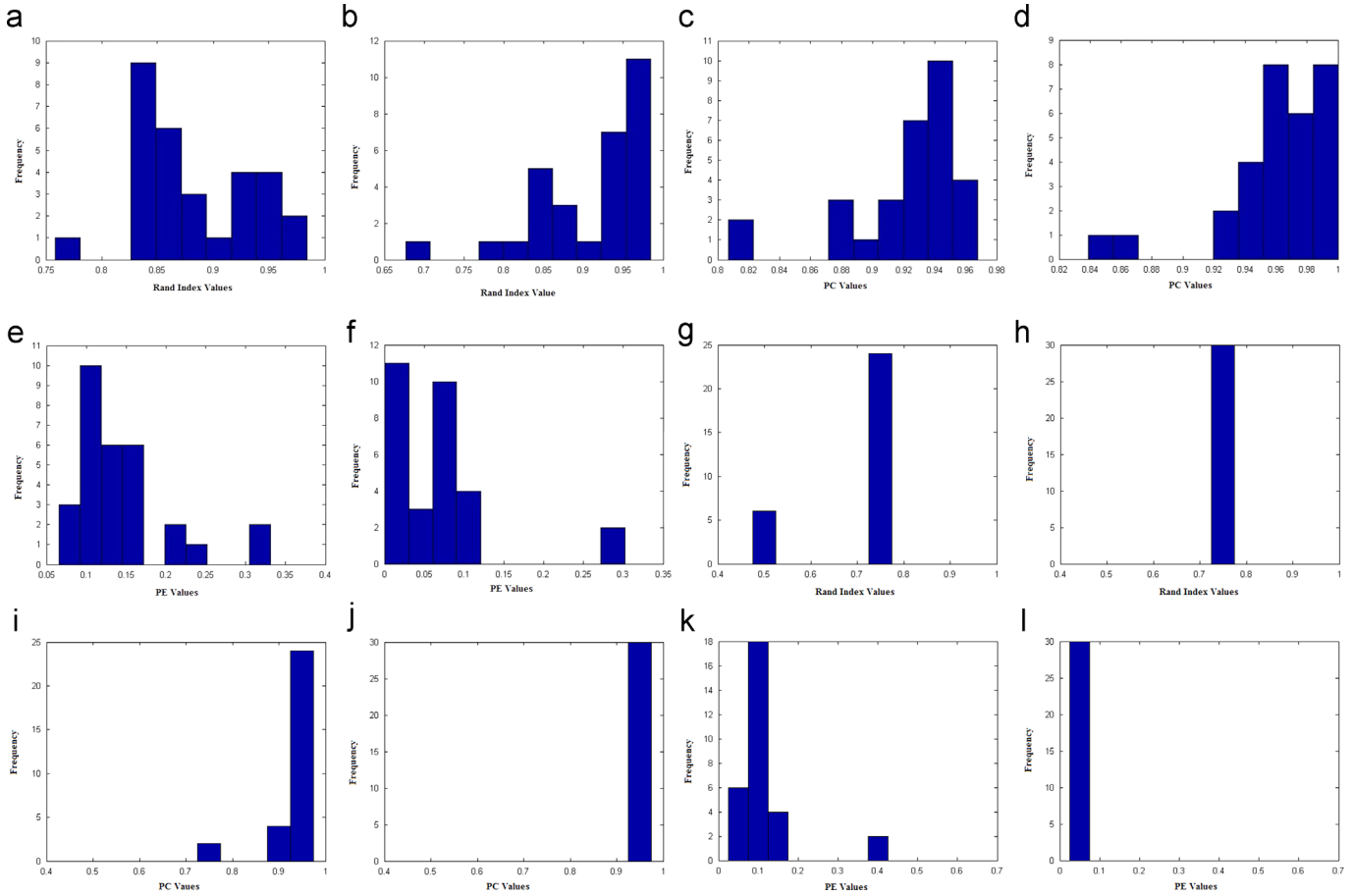

Fig. 7. For Zoo dataset, Rand Index corresponding to (a) $\beta=0$ (b) $\beta=9$; V_{pc}corresponding to (c) $\beta=0$ (d) $\beta=9$;V_{pe}corresponding to (e) $\beta=0$, (f) $\beta=9$. for Mushroom Dataset, Rand Index corresponding to (g) $\beta=0$ (h) $\beta=9$; V_{pc}corresponding to (i) $\beta=0$ (j) $\beta=9$; V_{pe}corresponding to (k) $\beta=0$ (l) $\beta=9$.

Total number of attribute values $=P=\sum_{l=1}^{p} m_l$, where m_l is the number of distinct attribute values possible for attribute l, $l=1,2,\ldots,p$.

The computational complexity of the 2nd, 3rd and 4th step of our algorithm are $O(pkn)$, $O(Pkn)$ and $O(pkn)$ respectively. Thus, total complexity of one iteration of our algorithm is of $O(kn(2p+P))$. Our algorithm requires $O(pn+pk+kn+p)$ storage space to hold the data points, the cluster centers, the membership matrix and the weights. The computational complexity in case of conventional fuzzy k-modes clustering algorithm [2] is given by $O(kn(m+M))$ and the algorithm requires $O(pn+pk+kn)$ storage space to hold the data points, the cluster centers and the membership matrix.

Though due to the additional step of computation of weights, our algorithm demands for more computational complexity than that of the conventional fuzzy k-modes algorithm, but the difference is asymptotically negligible when $P \gg p$, i.e. the total number of attribute values is much more than total number of attributes. The extra storage space required for storing the weight vector in our algorithm is negligible when the number of data points is very high.

5. Conclusion

A novel modification to the conventional fuzzy k-modes algorithm was proposed in this paper. Here, at each iteration, weights were adjusted to minimize the objective function and are used in determination of the cluster memberships. Detailed statistical analysis of the experimental results clearly shows that the new modification provides us with (statistically significant amount of) improvement over the fuzzy k-Modes algorithm, in terms of Rand Index and fuzziness (V_{pc} and V_{pe}). The proposed weighting scheme is quite general in nature and hence can be applied to any advanced variant of the fuzzy k-Modes algorithm.

Future works on this topic may include, finding some specific values of β, which works better than most other values of β, on a fairly large amount of datasets. The performance measures can be extended to a much larger set, in which aspects like geometric structure are taken care of.

Appendix

Here we present a proof of the fact that in the setup under consideration, the value of β can never be optimized. We shall give a proof of the following fact,

$$O^{m,\beta_1}(U_0,Z_0,X,W_0) \geq O^{m,\beta_2}(U_0,Z_0,X,W_0), \quad 0<\beta_1<\beta_2.$$

Proof:

$$O^{m,\beta}(U,Z,X,W)=\sum_{i=1}^{n} u_{ji}^{m} d_{ji}^{W}$$

$$=\sum_{i=1}^{n} u_{ji}^{m} \sum_{l=1}^{p} \delta^{W}(Z_{jl},X_{il})$$

$$= \sum_{i=1}^{n} u_{ji}^{m} \sum_{l=1}^{p} w_l^{\beta} \delta(z_{jl}, x_{il}) = \sum_{l=1}^{p} w_l^{\beta} \sum_{i=1}^{n} u_{ji}^{m} \delta(z_{jl}, x_{il})$$

$$O^{m,\beta_1}(U_0, Z_0, X, W_0) = \sum_{l=1}^{p} w_l^{\beta_1} \sum_{i=1}^{n} u_{ji}^{m} \delta(z_{jl}, x_{il})$$

$$O^{m,\beta_2}(U_0, Z_0, X, W_0) = \sum_{l=1}^{p} w_l^{\beta_2} \sum_{i=1}^{n} u_{ji}^{m} \delta(z_{jl}, x_{il})$$

Now we use the following fact

$f_c(x) = c^x,\ c \in [0, 1]$

Is a decreasing function. Under this notation

$$O^{m,\beta_1}(U_0, Z_0, X, W_0) = \sum_{l=1}^{p} f_{w_l}(\beta_1) \sum_{i=1}^{n} u_{ji}^{m} \delta(z_{jl}, x_{il})$$

$$O^{m,\beta_2}(U_0, Z_0, X, W_0) = \sum_{l=1}^{p} f_{w_l}(\beta_2) \sum_{i=1}^{n} u_{ji}^{m} \delta(z_{jl}, x_{il})$$

Now, a linear combination of two decreasing function with all coefficients non-negative is again decreasing. Thus, we are done.

We have proved that $O^{m,\beta}(U, Z, X, W)$ is a decreasing function of $\beta(\beta > 0)$ for fixed m, U, Z, X, W. So, we cannot optimize the cost function w.r.t. β.

Please check Appendix for Figs. A1. and A2.

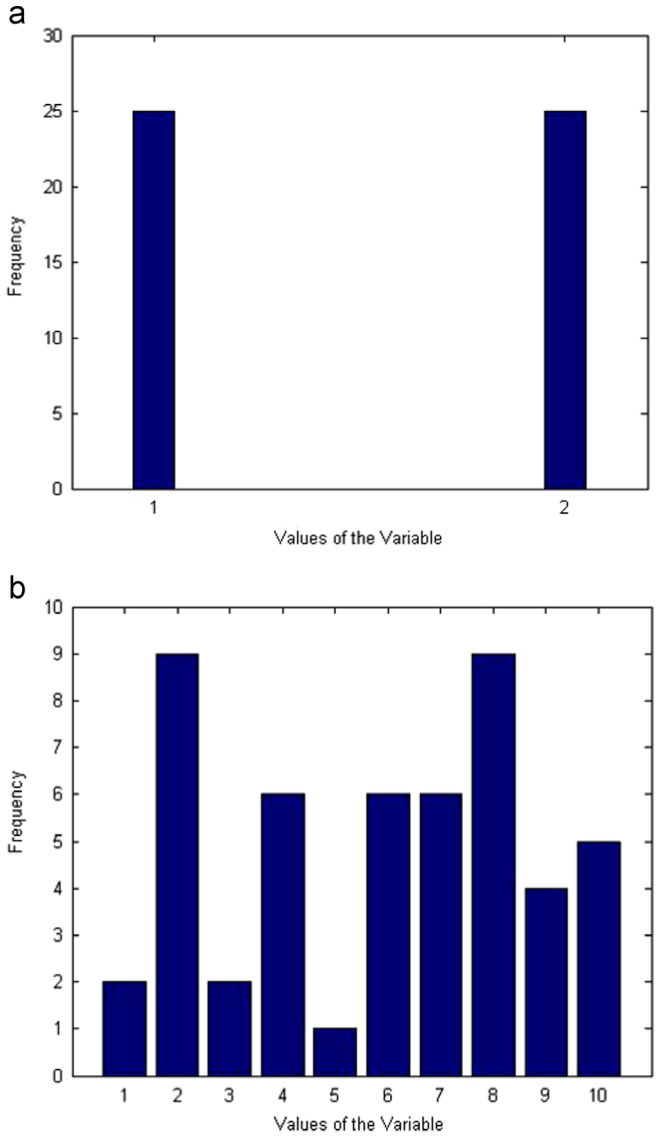

Fig. A1. For the synthetic data used, the plot of the frequencies of the values of the (a) first variable (main variable), and (b) second Variable (noise variable).

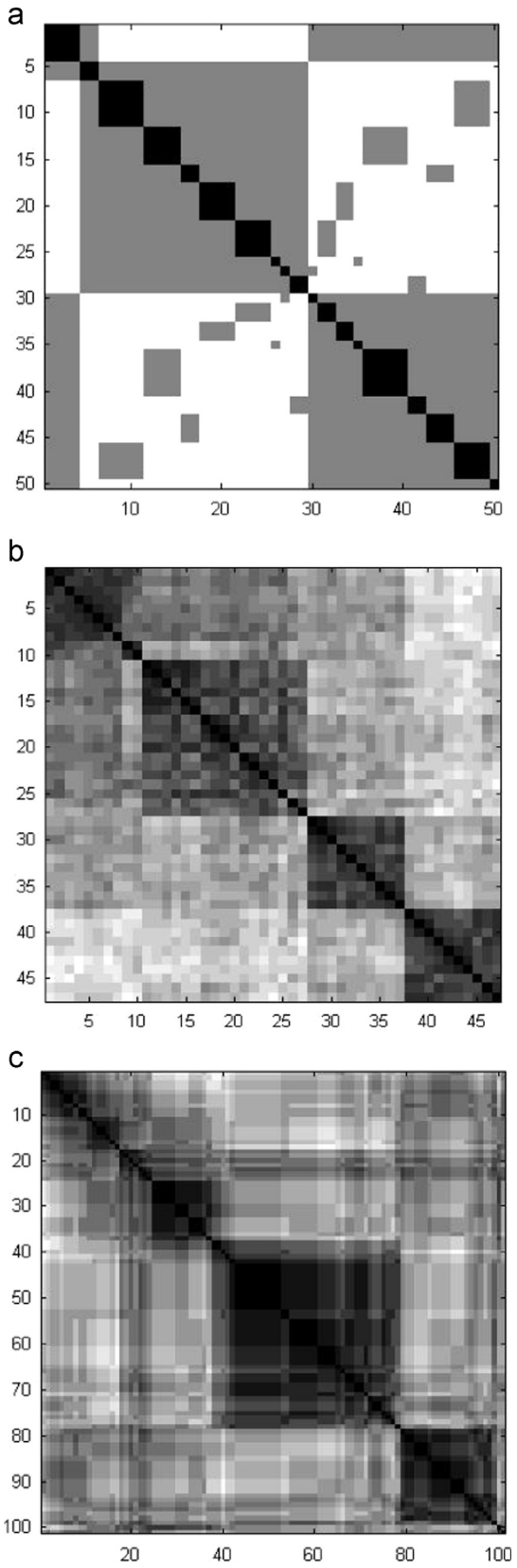

Fig. A2. The VAT pictures of (a) Synthetic (b) Soybean and (c) Zoo datasets.

References

[1] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666.

[2] Z. Huang, M.K. Ng, A fuzzy k-Modes algorithm for clustering categorical data, IEEE Trans. Fuzzy Syst. 7 (4) (1999).

[3] D.W. Kim, K.H. Lee, D. Lee, Fuzzy clustering of categorical data using fuzzy centroids, Pattern Recognit. Lett. 27 (5) (2005).

[4] Z. He, S. Deng, X. Xu, Improving k-modes algorithm considering frequencies of attribute values in mode, in: Proceedings of the International Conference on Computational Intelligence and Security, 2005, pp. 157–162.

[5] O. San, V. Huynh, Y. Nakamori, An alternative extension of the k-means algorithm for clustering categorical data, Int. J. Appl. Math Comput. Sci. 14 (2) (2004) 241–247.

[6] L. Bai, J. Liang, The k-modes type clustering plus between-cluster information for categorical data, Neurocomputing 133 (2014) 111–121.

[7] F. Cao, J. Liang, D. Li, X. Zhao, A weighting k-modes algorithm for subspace clustering of categorical data, Neurocomputing 108 (2013) 23–30.

[8] M.K. Ng, M.J. Li, J.Z. Huang, Z. He, On the impact of dissimilarity measures in k-modes clustering algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007).

[9] G. Gan, J. Wu, Z. Yang, A genetic fuzzy k-modes algorithm for clustering categorical data, Expert Syst. Appl. 36 (2009) 1615–1620.

[10] M.S. Yang, C.Y. Lin, Block fuzzy k-modes clustering algorithm, in: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Korea, August 20–24, 2009.

[11] L. Bai, J. Liang, C. Dang, F. Cao, A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets Syst. 215 (2013) 55–73.

[12] F. Cao, J. Liang, D. Li, C. Dang, A dissimilarity measure for the k-modes clustering algorithm, Knowl.-Based Syst. 26 (2012) 120–127.

[13] I. Saha, J.P. Sarkar, U. Maulik, Rough set based fuzzy k-modes for categorical data, in: B.K. Panigrahi et al. (Eds.), SEMCO, LNCS 7677, 2012, pp.323–330.

[14] O.S. Soliman, D.A. Saleh, S. Rashwan, A bio inspired fuzzy k-modes clustering algorithm, in: T. Huang et al. (Eds.), ICONIP, Part III, LNCS 7665, 2012, pp.663–669.

[15] W.S. Desarbo, J.D. Carroll, L.A. Clark, P.E. Green, Synthesized clustering: a method for amalgamating clustering bases with differential weighting variables, Psychometrika 49 (1984) 57–78.

[16] J.Z. Huang, M.K. Ng, H. Rong, Z. Li, Automated variable weighting in k-means type clustering, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005).

[17] C. Blake, E. Keogh, C.J. Merz, UCI Repository of Machine Learning Databases, Department of Information and Computer Sciences, University of California, Irvine, CA, 1998. ⟨http://www.ics.uci.edu/~mlearn/MLRepository.html⟩.

[18] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (336) (1971) 846–850.

[19] J.C. Bezdek, Cluster validity with fuzzy sets, J. Cybern. 3 (3) (1973) 58–73.

[20] J.C. Bezdek, Mathematical models for systematic and taxonomy, in: Proceedings of 8th International Conference on Numerical Taxonomy, San Francisco, 1975, pp. 143–166.

[21] J.C. Dunn, Indices of partition fuzziness and the detection of clusters in large data sets, in: M.M. Gupta (Ed.), Fuzzy Automata and Decision Process, Elsevier, New York, 1977.

[22] J.C. Bezdek, M. Windham, R. Ehrlich, Statistical parameters of fuzzy cluster validity functionals, Int. J. Comput. Inf. Sci. 9 (4) (1980) 323–336.

[23] S. Garcia, A. Fernandez, J. Luengo, F. Herrera, Advanced non-parametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power, Inf. Sci. 180 (10) (2010) 2044–2064.

[24] A. Keller, F. Klawonn., Fuzzy clustering with weighting of data variables, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 8 (2000) 735–746.

Arkajyoti Saha received his B. Stat (Hons.) degree from the Indian Statistical Institute, Kolkata, India, 2014. His research interest includes machine learning, Statistical pattern recognition, Fuzzy logic, Optimization Techniques. He has acted as a reviewer of peer-reviewed journals like Engineering Applications of Artificial Intelligence and Swarm and Evolutionary Computation. He is currently pursuing M. Stat degree at the Indian Statistical Institute, Kolkata, India.

Swagatam Das received his B.E. Tel. E., M.E. Tel. E (Control Engineering specialization) and Ph.D. degrees, all from the Jadavpur University, India, in 2003, 2005, and 2009 respectively. He served as an assistant professor at the department of Electronics and Telecommunication Engineering of Jadavpur University from 2006 to 2011.

Presently he is serving as an assistant professor at the Electronics and Communication Sciences Unit of Indian Statistical Institute, Kolkata. His current research interests include evolutionary computing, pattern recognition, multi-agent systems, and wireless communication. Dr. Das has published more than 150 research articles in peer-reviewed journals and international conferences. He is the founding co-editor-in-chief of "Swarm and Evolutionary Computation", an international journal from Elsevier. He serves as associate editors of the IEEE Trans. on Systems, Man, and Cybernetics: Systems, IEEE Computational Intelligence Magazine, IEEE Access, Neurocomputing, Engineering Applications of Artificial Intelligence, and Information Sciences (Elsevier). He is an editorial board member of Progress in Artificial Intelligence (Springer), International Journal of Artificial Intelligence and Soft Computing and International Journal of Adaptive and Autonomous Communication Systems. He has been acting as a regular reviewer for journals like Pattern Recognition, IEEE Transactions on Evolutionary Computation, IEEE/ACM Transactions on Computational Biology and Bioinformatics, IEEE Transactions on SMC Part A, Part B, and Part C. He has acted as guest editors for special issues in journals like IEEE Transactions on Evolutionary Computation and IEEE Transactions on SMC, Part C. He co-authored a research monograph on metaheuristic clustering techniques from Springer in 2009.