Robust and Sparse Fuzzy K-Means Clustering

Jinglin Xu, Junwei Han*, Kai Xiong, Feiping Nie*
Northwestern Polytechnical University, Xi'an, 710072, P. R. China
xujinglinlove, junweihan2010, bearkai1992, feipingnie@gmail.com

Abstract

The partition-based clustering algorithms, like K-Means and fuzzy K-Means, are most widely and successfully used in data mining in the past decades. In this paper, we present a robust and sparse fuzzy K-Means clustering algorithm, an extension to the standard fuzzy K-Means algorithm by incorporating a robust function, rather than the square data fitting term, to handle outliers. More importantly, combined with the concept of sparseness, the new algorithm further introduces a penalty term to make the object-clusters membership of each sample have suitable sparseness. Experimental results on benchmark datasets demonstrate that the proposed algorithm not only can ensure the robustness of such soft clustering algorithm in real world applications, but also can avoid the performance degradation by considering the membership sparsity.

1 Introduction

The classical K-Means problem is a clustering algorithm which assigns a set of data points into clusters so that the data points in the same cluster have high similarity but are dissimilar if they belong to other clusters. The K-Means algorithm is widely used due to its efficiency. A variety of modifications and generalizations have been proposed and developed over the years. Among different variants of K-Means algorithm, fuzzy K-Means (FKM) algorithm is the most popular. It was originally proposed by [Ruspini, 1969] and had been modified by [Bezdek, 1980]. The essential difference between K-Means and FKM algorithms is FKM allows a data point to have memberships in all clusters rather than having a distinct membership to one single cluster. The K-Means problem is a well-known example for a hard clustering, whereas FKM is a continuous generalization of the K-Means problem that is named as a soft clustering. For the reason that ambiguity exists in real world datasets, FKM clustering has gained more attention recently.

It can be seen that these K-Means-type (hard or soft) algorithms can effectively tackle numerous problems in various fields such as medical imaging, target recognition and image segmentation. Because the philosophy of K-Means-type algorithms is extensively used, a large number of modifications had been proposed [Feiping *et al.*, 2011; Cai *et al.*, 2013]. Here we only discuss some of typical algorithms related to the proposed method.

There are two major issues in the application of FKM algorithms. The first issue is the lack of robustness during clustering. Generally FKM algorithms use a Euclidean distance measure to assign memberships to each sample for clustering, which only can provide good clustering result without outliers. To overcome this drawback, the work [Zhang *et al.*, 2003] replaced the Euclidean norm with kernel distance measures. However, this method does not consider any spatial dependence of the data elements, which not only makes it very sensitive to outliers but also takes more time to converge the algorithm. Recently, a set of novel FKM algorithms have been formulated to address this issue by using new techniques and improve their performance. For instance, [Ji *et al.*, 2011] introduced the local spatial weights in the objective function, which allows the suppression of outliers and helps to resolve ambiguity. [Zhao *et al.*, 2011] introduced a non-local spatial constraint term into the objective function to deal with image noise more effectively. [Ji *et al.*, 2012] combined local spatial information embedded in the data to further improve its robustness to outliers. [Kannan *et al.*, 2012] proposed robust FKM based kernel function by incorporating normed kernel function and center initialization algorithm. [Wang *et al.*, 2013] incorporated an adaptive spatial information-theoretic fuzzy clustering into the conventional FKM to improve the robustness.

The second issue is the difficulty of choosing an appropriate regularization for the FKM algorithm. It is well-known that FKM can be extended with regularization to reduce the effect of outliers and further improve its performance. In [Li *et al.*, 2008; Namkoong *et al.*, 2010], the problem can obtain the desired solution by:

$$\min_{\mathbf{x}} \|\mathbf{x}-\mathbf{y}\| + \gamma\Phi(\mathbf{x})$$

where $\mathbf{x}$ is the variable to be solved, $\mathbf{y}$ is the variable observed, and $\gamma(>0)$ is a parameter of regularization. $\Phi(\mathbf{x})$ is used to restrict the admissible solution within the space of smooth functions. However, there is no rule for selecting the $\Phi(\mathbf{x})$. In general, the derivative of $\mathbf{x}$ is chosen for $\Phi(\mathbf{x})$,

*Corresponding authors

which measures the roughness of $\mathbf{x}$. Recently, regularization has been applied to the clustering problem and several objective functions were formulated using different regularization terms such as [Li and Mukaidono, 1995; Miyamoto and Umayahara, 1998; Özdemir and Akarun, 2002; Yu and Yang, 2007]. Although these regularization-based methods are better than previous methods, deciding regularization constant is an important problem which was set empirically and time consuming.

On the basis of the above analysis, it can be seen that FKM is a simple and effective method, however, its membership values might be inaccurate in an outlier environment. One of the reasons is the distance measures without robustness used in clustering. Another reason is the distributions and sparseness on memberships. To address these two weaknesses, we propose a robust and sparse fuzzy K-Means clustering algorithm by incorporating robust loss function, rather than the square data fitting term, to handle outliers, and further combining the concept of sparsity, which introduces a regularization, to make their memberships of each sample with respect to different clusters have suitable sparseness.

This paper tends to address the above highlighted drawbacks by introducing a novel robust objective function of fuzzy K-Means with a regularization about the concept of sparsity. Furthermore, the dynamic γ strategy is given in this paper, which suggests that the proposed method can be improved further if the parameter γ can be estimated correctly. A large number of experiments demonstrate that the proposed method is more powerful in clustering benchmark datasets.

2 Related Work

There is a large number of KM extensions proposed in past years, such as [Pham, 2001; Stelios and Vassilios, 2010; Nie *et al.*, 2014a]. Due to the limited space, we only review some closely relevant work as follows.

2.1 Fuzzy C-Means Algorithm

FCM is one of the most popular fuzzy clustering techniques, which was proposed by Dunn [Dunn, 1973] and eventually modified by Bezdek [Bezdek, 1980]. In this approach, the data points have their membership values with the cluster centers, which will be updated iteratively.

Let $\mathbf{X}=[\mathbf{x}_1,...,\mathbf{x}_n]$ be a set of n objects. To cluster $\mathbf{X}$ into c clusters, the standard fuzzy C-Means algorithm minimizes the following objective function:

$$f[\mathbf{U},\mathbf{V}]=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}\|\mathbf{x}_i-\mathbf{v}_k\|^2 \tag{1}$$

where f is called the loss function, m is an appropriate level of cluster fuzziness, $\mathbf{v}_k$ is interpreted as the centroid of the k-th cluster, and u_{ik} denotes the grade of membership of the i-th object in the k-th cluster and satisfies the following conditions:

$$u_{ik}\in[0,1], 1\le i\le n, 1\le k\le c$$
$$\sum_{k=1}^{c}u_{ik}=1, \sum_{i=1}^{n}u_{ik}>0$$

To minimize (1) subject to $\sum_{k=1}^{c}u_{ik}=1$ by using Lagrangian multiplier method, a considered point was demonstrated to be a local minimum solution of (1) if and only if:

$$u_{ik}=\frac{1}{\sum_{s=1}^{c}(\frac{d_{ik}}{d_{is}})^{\frac{2}{m-1}}} \tag{2}$$

$$\mathbf{v}_k=\frac{\sum_{i=1}^{n}u_{ik}^{m}\mathbf{x}_i}{\sum_{i=1}^{n}u_{ik}^{m}} \tag{3}$$

where $d_{ik}=\|\mathbf{x}_i-\mathbf{v}_k\|^2$.

The iterative FCM algorithm is stopped if $max|u_{ik}^{(t+1)}-u_{ik}^{(t)}|<\varepsilon$ where ε is a small positive integer and t denotes number of iterations [Bezdek, 2013; Isazadeh and Ghorbani, 2003]. Noted that for $m=1$, FCM algorithm converges in theory to the traditional K-Means solution [Smyth, 2000].

2.2 Agglomerative Fuzzy K-Means Algorithm

To tackle some issues during clustering, like the number of clusters and initial cluster centers, [Li *et al.*, 2008] introduced a penalty term to the objective function of FKM. Clustering $\mathbf{X}$ into c clusters by this algorithm is to minimize the following objective function:

$$f[\mathbf{U},\mathbf{V}]=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}d_{ik}+\gamma\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\log u_{ik} \tag{4}$$

subject to

$$\sum_{k=1}^{c}u_{ik}=1, u_{ik}\in(0,1], 1\le i\le n, 1\le k\le c$$

where $\mathbf{U}$ is an n-by-c partition matrix, $\mathbf{V}$ is an c-by-m matrix containing the cluster centers, and d_{ik} is a dissimilarity measure between the k-th cluster center and the i-th object.

The alternating minimization procedure between membership matrix $\mathbf{U}$ and cluster center matrix $\mathbf{V}$ can be applied to (4), which follows:

$$u_{ik}=\frac{\exp(\frac{-d_{ik}}{\gamma})}{\sum_{s=1}^{c}\exp(\frac{-d_{is}}{\gamma})} \tag{5}$$

$$\mathbf{v}_k=\frac{\sum_{i=1}^{n}u_{ik}\mathbf{x}_i}{\sum_{i=1}^{n}u_{ik}} \tag{6}$$

The first term in (4) is the cost function of the standard K-Means algorithm. The second term is added to maximize the negative objects-to-clusters membership entropy in the clustering process, which can simultaneously minimize the within cluster dispersion and maximize the negative weight entropy to determine clusters to contribute to the association of objects.

3 The Proposed Method

3.1 Formulation

To overcome the outlier sensitivity problem of the data-driven term and achieve an optimal approximate solution in related work, we formulate a robust and sparse fuzzy K-Means clustering combined with different kinds of norms, like $\ell_{2,1}$-norm

[Ding *et al.*, 2006] and capped ℓ_1-norm [Jiang *et al.*, 2015], to make the proposed algorithm more robust to outliers. It is to minimize the following objective function:

$$f[\mathbf{U},\mathbf{V}]=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\tilde{d}_{ik}+\gamma\|\mathbf{U}\|_F^2 \tag{7}$$

subject to

$$\mathbf{U1}=\mathbf{1},\mathbf{U}\geq 0$$

where $\mathbf{U}$ is an n-by-c membership matrix. $\tilde{d}_{ik}$ is a measure which can be flexibly defined as some alternative norms to measure the dissimilarity by different ways. For instance, $\tilde{d}_{ik}=\|\mathbf{x}_i-\mathbf{v}_k\|_2^2$ is commonly used in clustering. In this paper, we replace this *Least Square* term with *Least Absolute* term $\tilde{d}_{ik}=\|\mathbf{x}_i-\mathbf{v}_k\|_2$ ($\ell_{2,1}$-norm) and the capped ℓ_1-norm term $\tilde{d}_{ik}=\min(\|\mathbf{x}_i-\mathbf{v}_k\|_2,\varepsilon)$ (where ε is a threshold), where they can be more robust to outliers than Least Square.

3.2 The Properties of the Algorithm

Fuzziness

According to [Bauckhage, 2015], traditional K-Means clustering was rigorously established where the membership matrix $\mathbf{U}$ is binary such that each row of $\mathbf{U}$ contains a single 1 and $c-1$ elements that are 0. The rows of $\mathbf{U}$ sum up to 1 and its column sums indicate the number elements per cluster.

Differed from hard K-Means clustering, the proposed method relaxes each element of $\mathbf{U}$ into a nonnegative value less than 1 under the constraint conditions. Note that the proposed method is different from fuzzy C-Means clustering, because the latter also requires to choose an appropriate level of cluster fuzziness $m\geq 1$ (mostly $m=2$), and the former presets $m=1$.

Robustness

It is well known that the quadratic loss function is not robust to outliers. To overcome this weakness, the quadratic loss function should be replaced by an insensitive one to outliers, e.g. $\ell_{2,1}$-norm and capped ℓ_1-norm. In this paper, we use the objective functions based on $\ell_{2,1}$-norm and capped ℓ_1-norm, respectively. Based on $\ell_{2,1}$-norm which is not squared and usually used to induce sparsity, outliers have less importance than the squared one. Based on the capped ℓ_1-norm, the loss function treats $\mathbf{r}$ equally if $\|\mathbf{r}\|_2$ is bigger than ε, which is more robust to outliers than the squared one.

Sparseness

During clustering, the proposed algorithm tries to minimize the robust residual term and consider the sparsity of memberships of each object being assigned to different clusters simultaneously. The importance of each part in the minimization process of (7) is balanced by the parameter γ.

It can be seen that the square of memberships of each object measures whether the object is assigned to a single cluster or several clusters. In the case of assigning to a single cluster, $\sum_{k=1}^{c}u_{ik}^2$ is equal to 1. In the case of assigning to several clusters, $\sum_{k=1}^{c}u_{ik}^2$ is a positive number and much smaller than 1. Minimization of the sum of memberships of each object is to assign each object to more clusters instead of a single cluster.

If the parameter γ is zero, the membership vector of each object is sparse (Only one element is nonzero and others are zero). When γ is greater than zero, the membership vectors are less sparse than that in case of γ being zero. The sparseness of the membership vectors is a progressive change when we tune γ. Along with the gradual increase of γ, membership vectors contain a growing number of nonzero elements. When γ reaches up to a large value, all elements of membership vectors are nonzero, and membership vectors are non-sparse at this time. Thus, the parameter γ controls the sparseness degree of the membership vectors. One of our goals is to find the reasonable sparseness of membership vectors to obtain more accurate clustering results.

3.3 The Optimization Procedure

In this paper, we want to find a set of highly accurate centroids to better group objects. An straightforward way is to use *Least Square* loss function. However, to provide better robustness, we go further to use $\ell_{2,1}$-norm and capped ℓ_1-norm loss functions, respectively. Concretely, the objective values of capped ℓ_1-norm loss dose not increase any more if $\|\bullet\|_2$ is larger than ε. Therefore, $\ell_{2,1}$-norm loss is more robust than *Least Square* loss, but might be less robust than capped ℓ_1-norm. Thus, the objective function of our robust and sparse fuzzy K-Means clustering algorithms are formulated using different robust norms ($\ell_{2,1}$-norm and capped ℓ_1-norm) as follows:

$$\min_{\mathbf{U1}=\mathbf{1},\mathbf{U}\geq 0,\mathbf{V}}\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\|\mathbf{x}_i-\mathbf{v}_k\|_2+\gamma\|\mathbf{U}\|_F^2 \tag{8}$$

$$\min_{\mathbf{U1}=\mathbf{1},\mathbf{U}\geq 0,\mathbf{V}}\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\min(\|\mathbf{x}_i-\mathbf{v}_k\|_2,\varepsilon)+\gamma\|\mathbf{U}\|_F^2 \tag{9}$$

Although the norms are different to each other, both of them can be optimized by using the iterative re-weighted method proposed in [Nie *et al.*, 2010; 2014b]. These two methods ((8) and (9)) extend FKM with robust norms and regularizations to reduce the effect of outliers and keep the memberships with proper sparsity.

Concretely, $\mathbf{U}$, $\mathbf{V}$ and auxiliary variable s_{ik} are updated by following updating rules:

$$\min_{\mathbf{U1}=\mathbf{1},\mathbf{U}\geq 0,\mathbf{V}}\sum_{i=1}^{n}\sum_{k=1}^{c}s_{ik}u_{ik}\|\mathbf{x}_i-\mathbf{v}_k\|_2^2+\gamma\|\mathbf{U}\|_F^2 \tag{10}$$

For the (8), auxiliary variable s_{ik} is defined as $\frac{1}{2\|\mathbf{x}_i-\mathbf{v}_k\|_2}$. For the (9), auxiliary variable s_{ik} is defined as follows:

$$s_{ik}=\begin{cases}\frac{1}{2\|\mathbf{x}_i-\mathbf{v}_k\|_2}, & \|\mathbf{x}_i-\mathbf{v}_k\|_2\leq\varepsilon\\ 0, & \|\mathbf{x}_i-\mathbf{v}_k\|_2>\varepsilon\end{cases} \tag{11}$$

From two definitions of s_{ik}, it is obvious that when a sample is not outlier, namely, $\|\mathbf{x}_i-\mathbf{v}_k\|_2\leq\varepsilon$, this sample with lower reconstruction error has higher weight. At this moment, our two methods are equivalent. When a sample is a outlier, namely $\|\mathbf{x}_i-\mathbf{v}_k\|_2>\varepsilon$, this sample with higher reconstruction error has lower weight. Nevertheless, we expect to enhance the outlier insensitiveness of our method, and even hope the

weight of outlier to be 0. Thus, to provide better robustness, we go further to solve the problem of (9).

The (10) is convex separately with respect to $\mathbf{U}$ and $\mathbf{V}$, thus we solve it by updating $\mathbf{U}$ and $\mathbf{V}$ alternately.

With $\mathbf{V}$ fixed, the objective function becomes:

$$\min_{\mathbf{U1}=\mathbf{1},\mathbf{U}\geq 0}\sum_{i=1}^{n}\sum_{k=1}^{c}(s_{ik}u_{ik}\|\mathbf{x}_i-\mathbf{v}_k\|_2^2+\gamma u_{ik}^2) \tag{12}$$

For each $\mathbf{x}_i$, (12) can be separated into n subproblems:

$$\min_{\mathbf{u}^i\mathbf{1}=1,\mathbf{u}^i\geq 0}\sum_{k=1}^{c}(h_{ik}u_{ik}+\gamma u_{ik}^2) \tag{13}$$

where $\mathbf{u}^i$ is the i-th line of matrix $\mathbf{U}$ and $h_{ik}=s_{ik}\|\mathbf{x}_i-\mathbf{v}_k\|_2^2$ is an element of matrix $\mathbf{H}$. After being simplified, (13) can be written as:

$$\min_{\substack{\mathbf{u}^i\mathbf{1}=1,\\ \mathbf{u}^i\geq 0}}\|\mathbf{u}^i-\tilde{\boldsymbol{h}}^i\|_2^2 \tag{14}$$

where $\mathbf{u}^i$ is a variable to be optimized and the row vector $\tilde{\boldsymbol{h}}^i=\frac{-\mathbf{h}^i}{2\gamma}$ is a constant in this stage. We utilize the technique of [Huang *et al.*, 2015] to solve (14) that updates the membership vector. It is the fact that the solution of (14) must be sparse.

With $\mathbf{U}$ fixed, the objective of (10) becomes:

$$\min_{\mathbf{V}}\sum_{i=1}^{n}\sum_{k=1}^{c}s_{ik}u_{ik}\|\mathbf{x}_i-\mathbf{v}_k\|_2^2 \tag{15}$$

which can be decomposed into c independent problems as follows:

$$\min_{\mathbf{v}_k}\sum_{i=1}^{n}s_{ik}u_{ik}\|\mathbf{x}_i-\mathbf{v}_k\|_2^2 \tag{16}$$

Each iteration of (16) involves minimizing a quadratic objective function. Using the Lagrange multiplier method, the global optimum can be reached by taking derivatives and setting them to zeros. Thus, there is:

$$\mathbf{v}_k=\frac{\sum_{i=1}^{n}s_{ik}u_{ik}\mathbf{x}_i}{\sum_{i=1}^{n}s_{ik}u_{ik}} \tag{17}$$

In addition, assuming that $\mathbf{U}^{(t)}$ and $\mathbf{V}^{(t)}$ are computed from the solution of the t-th iteration, we can update the nonnegative auxiliary variable s_{ik} according to (11) by the current $\mathbf{V}^{(t)}$ when we solve (9). When we solve (8), we can update s_{ik} according to $\frac{1}{2\|\mathbf{x}_i-\mathbf{v}_k\|_2}$ with the current $\mathbf{V}^{(t)}$.

The whole algorithm of our method is listed in Algorithm 1. Because the proposed RSFKM meets the conditions of the references [Nie *et al.*, 2010; 2014b], it can be easily proved that Algorithm 1 is absolutely converged.

Algorithm 1 The algorithm of RSFKM method

Input:

Data matrix $\mathbf{X}$, the number of clusters c, regularization parameter γ and threshold ε.

Output:

Clustering indicator matrix $\mathbf{U}$ and Cluster centroid matrix $\mathbf{V}$.

Initialization:

Set $t=0$. Initialize $\mathbf{U}$, $\mathbf{V}$ and auxiliary variable s by $\mathbf{U1}=\mathbf{1}$, $\mathbf{U}\geq 0$, and $s_{ik}=1$ for $i=1,\ldots,n$; $k=1,\ldots,c$.

While *not converge* **do**

1: Solve $\mathbf{U}$ by (14);
2: Update $\mathbf{V}$ by (17);
3: Update s_{ik} by (11) if solve (9). Update s_{ik} using $\frac{1}{2\|\mathbf{x}_i-\mathbf{v}_k\|_2}$ if solve (8).

End While, return $\mathbf{U}$, $\mathbf{V}$.

4 Experiments

4.1 Datasets

We evaluate the performance of the proposed method (RSFKM) on three benchmark datasets in terms of two standard clustering evaluation metrics, namely, Accuracy (ACC) and Normalized Mutual Information (NMI) [Cai *et al.*, 2005]. Among those, two datasets are image datasets, COIL-20[1] and COIL-100[2]. The rest is the MNIST[3] database of handwritten digits. We provide here four smaller subsets (MINIST-2K2K, MINIST-10K, MINIST-TEST and MINIST-ORIG) from MINIST. Table 1 summarizes the characteristics of these datasets used in our experiments and Figure 1 shows some example images from different datasets.

Table 1: Descriptions of benchmark datasets.

Dataset	♯ Samples	♯ Dimensions	♯ Classes
COIL-20	1440	60	20
COIL-100	7200	160	100
MINIST-2K2K	4000	120	10
MINIST-10K	10000	120	10
MINIST-TEST	10000	115	10
MINIST-ORIG	70000	120	10

4.2 Experiment Setup

Comparison Methods

We evaluate the performance of the proposed method on benchmark datasets. We compare our method RSFKM (with $\ell_{2,1}$-norm and capped ℓ_1-norm) with K-means (KM), fuzzy C-Means (FCM, $m>1$), agglomerative fuzzy K-Means (AFKM), sparse fuzzy K-Means (SFKM). Concretely, KM and FCM are classic clustering methods. The difference of AFKM and SFKM is that they are derived from different cluster fuzziness m and different regularization terms. We compare with SFKM due to two reasons. First, it has close relation with the proposed method. Second, the comparison with SFKM can show the advantage of robustness in our method.

Parameter Discussion

There are two important parameters in RSFKM, namely regularization parameter γ and threshold value ε. Each of them

[1] http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

[2] http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php

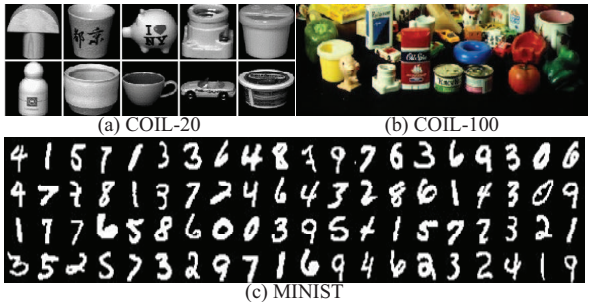
[3] http://yann.lecun.com/exdb/mnist/

Table 2: *ACC*(%±std) of different methods on different benchmark datasets.

Method	KM	FCM	AFKM	SFKM	RSFKM ($\ell_{2,1}$)	RSFKM (cappedℓ_1)
COIL-20	59.71 ± 4.59	63.96 ± 1.80	59.08 ± 2.88	65.03 ± 3.32	66.63 ± 2.63	**67.06±2.87**
COIL-100	47.39 ± 1.99	48.91 ± 1.29	41.64 ± 1.19	48.91 ± 1.84	49.82 ± 2.48	**51.00±1.16**
MINIST-2K2K	50.72 ± 3.05	49.55 ± 0.82	49.12 ± 5.47	52.53 ± 2.14	52.96 ± 2.30	**53.22±2.81**
MINIST-10K	54.45 ± 3.44	54.15 ± 0.83	52.54 ± 3.97	56.20 ± 3.81	**57.32±5.80**	56.51 ± 3.91
MINIST-TEST	55.84 ± 2.97	55.32 ± 1.72	52.14 ± 2.82	55.51 ± 3.43	55.51 ± 2.08	**56.72±2.76**
MINIST-ORIG	55.43 ± 4.26	54.80 ± 1.86	50.52 ± 3.53	56.83 ± 3.03	57.17 ± 3.44	**57.57±2.64**

Table 3: *NMI*(%±std) of different methods on different benchmark datasets.

Method	KM	FCM	AFKM	SFKM	RSFKM ($\ell_{2,1}$)	RSFKM (cappedℓ_1)
COIL-20	75.55 ± 1.73	74.05 ± 0.79	74.82 ± 1.95	76.38 ± 1.96	76.50 ± 1.75	**76.54±1.35**
COIL-100	76.65 ± 0.65	77.11 ± 0.40	73.82 ± 0.94	77.27 ± 0.50	77.62 ± 0.56	**77.70±0.39**
MINIST-2K2K	46.10 ± 1.54	44.25 ± 1.04	41.92 ± 3.33	47.32 ± 1.16	48.10 ± 1.81	**48.38±0.94**
MINIST-10K	50.82 ± 1.50	49.00 ± 0.92	45.93 ± 2.91	51.52 ± 2.18	**52.52±1.34**	52.27 ± 1.25
MINIST-TEST	51.97 ± 1.14	50.78 ± 0.95	46.11 ± 1.65	52.60 ± 1.98	52.92 ± 1.69	**53.17±1.42**
MINIST-ORIG	50.46 ± 1.94	48.92 ± 0.79	43.29 ± 1.34	51.38 ± 1.61	52.23 ± 1.23	**52.42±1.12**



(a) COIL-20 (b) COIL-100

(c) MINIST

Figure 1: Some example images from (a) COIL-20, (b) COIL-100 and (c) MINIST datasets.

plays an important role in each algorithm and should be determined carefully.

For the regularization parameter γ, it puts a restriction on the minimum distance between a data point and a cluster center and prevents membership from having extreme values, 0 and 1. Large γ makes the regularization term to dominate the objective function, and thus makes u_{ik} to be approximately equal to $\frac{1}{c}$. Small γ makes the residual term to dominate the objective function, and thus makes u_{ik} to be sparse. The value of γ, therefore, should be chosen carefully to balance the residual term and the regularization term. In this paper, the optimal value of γ was set empirically using the grid search method in a range from $[10^{-1}, 10^{1}]$ every 0.5 step.

For the threshold value ε, it mainly controls the number of outliers and is related to the residuals of representations. If the residual of a sample to centroid is larger than ε, it is regarded as outlier and not used to learn centroid matrix $\mathbf{V}$ since the corresponding s_{ik} is zero. If the residual of a sample to centroid is less than ε, s_{ik} is nonzero. In order to minimize the objective function $u_{ik}s_{ik}\|\mathbf{x}_i - \mathbf{v}_k\|_2^2 + u_{ik}^2$, u_{ik} should be small, and there will be two extreme cases. The first case is that one membership value of the sample $\mathbf{x}_i$ tends to 1 and others tend to 0, which makes the membership values too sparse and finally degrades into a hard clustering. The second case is that all memberships of the sample $\mathbf{x}_i$ become equally and their sum is 1, which makes the memberships not be sparse and finally degrades into the traditional FKM clustering. In order to avoid above two extreme cases, we need a tradeoff between the regularization parameter γ and the threshold value ε in capped ℓ_1-norm. Here we select ε in a range of $[0, 3]$.

In our experiments, we tune γ and ε appropriately using the grid search method based on different datasets. For different sets of (γ, ε), we calculate average ACC and NMI by repeat clustering 10 times, and then report the best result for each method, respectively.

Intuitively, in Figure 2, we show some memberships of one sample on four methods KM, FCM, AFKM and RSFKM (with capped ℓ_1-norm) to demonstrate the appropriate sparseness of the proposed method.

4.3 Experiment Results

Table 2 and 3 summarize the results of all methods on the benchmark datasets. We bold the corresponding results if they are significant better than results from other methods. It can be observed that the proposed method (RSFKM) outperforms other methods on all datasets according to metrics of ACC and NMI.

In particular, our method RSFKM (whether with $\ell_{2,1}$-norm or with capped ℓ_1-norm) significantly gets a better result than K-Means, FCM, AFKM and SFKM on all datasets. It is obvious that compared with SFKM and RSFKM, both K-Means and FCM cannot obtain good performance during clustering, due to the lack of robust and sparse information. In addition, our method RSFKM (with $\ell_{2,1}$-norm or capped ℓ_1-norm) can achieve more robustness than SFKM on almost all datasets. Meanwhile we observe that SFKM has done much better than K-Means and FCM. It can be seen that adding the regularization term into the objective function is essentially necessary. Finally, although RSFKMs (with $\ell_{2,1}$-norm and capped ℓ_1-norm) are all proposed by this paper, the robustness of them is different and usually depends on the different datasets. Note that the latter could get better results with proper parameters

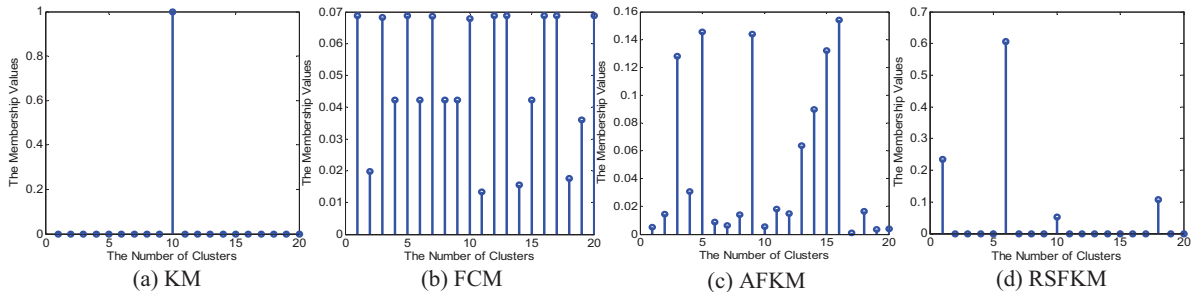


Figure 2: The membership values of each sample for three methods (KM, FCM, AFKM and RSFKM) on COIL-20 dataset.

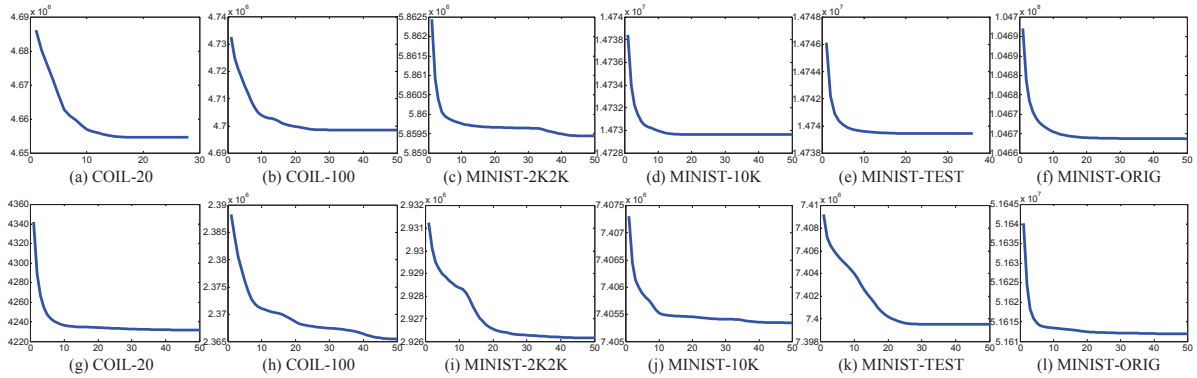


Figure 3: The convergence curves of RSFKM on COIL-20, COIL-100 and MINIST datasets.

(regularization parameter γ and the threshold ε of the capped ℓ_1-norm).

In a word, the experimental results in Table 2 and 3 demonstrate that our method taking into account both the robustness of the residual term and the sparseness of membership values for each sample can achieve better performance comparing with other state-of-the-art clustering approaches.

Furthermore, we test the convergence of RSFKM on COIL-20, COIL-100 and MINIST datasets. The results are shown in Figure 3. (a)-(f) denote the convergence curves of RSFKM with capped ℓ_1-norm. (g)-(l) denote the convergence curves of RSFKM with $\ell_{2,1}$-norm. It is can be seen that RSFKM algorithm can absolutely converge with few iteration steps.

5 Conclusion

In this paper, we have proposed a novel method, called robust and sparse fuzzy K-Means clustering algorithm, to obtain a more accurate clustering result. The proposed method minimized the objective function to deal with the effect of outliers considering sparse membership values by a re-weighted method, which is the weighted sum of the fuzzy K-Means with robust norms ($\ell_{2,1}$-norm and capped ℓ_1-norm) and the sparse quadratic regularization. The effectiveness of our method was demonstrated by a number of experiments on three benchmark datasets. In addition, determination of the regularization parameter is an important problem, and we analyzed its change strategy and set it empirically under investigation.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grants 61522207 and 61473231.

References

[Bauckhage, 2015] Christian Bauckhage. k-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548*, 2015.

[Bezdek, 1980] James C Bezdek. A convergence theorem for the fuzzy isodata clustering algorithms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):1–8, 1980.

[Bezdek, 2013] James C Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[Cai *et al.*, 2005] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *Knowledge and Data Engineering, IEEE Transactions on*, 17(12):1624–1637, 2005.

[Cai *et al.*, 2013] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *International Joint Conference on Artificial Intelligence*, 2013.

[Ding *et al.*, 2006] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R 1 -pca: rotational invariant l 1 -norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, pages 281–288, 2006.

[Dunn, 1973] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

[Feiping *et al.*, 2011] Nie Feiping, Zeng Zinan, Ivor W Tsang, Xu Dong, and Zhang Changshui. Spectral embedded clustering: a framework for in-sample and out-of-sample spectral clustering. *IEEE Transactions on Neural Networks*, 22(11):1796–1808, 2011.

[Huang *et al.*, 2015] Jin Huang, Feiping Nie, and Heng Huang. A new simplex sparse learning model to measure data similarity for clustering. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3569–3575. AAAI Press, 2015.

[Isazadeh and Ghorbani, 2003] A Isazadeh and M Ghorbani. Fuzzy c-means and its generalization by lp norm space. *WSEAS Transactions Mathematics*, 2(1):168–170, 2003.

[Ji *et al.*, 2011] Ze-Xuan Ji, Quan-Sen Sun, and De-Shen Xia. A modified possibilistic fuzzy c-means clustering algorithm for bias field estimation and segmentation of brain mr image. *Computerized Medical Imaging and Graphics*, 35(5):383–397, 2011.

[Ji *et al.*, 2012] Zexuan Ji, Yong Xia, Qiang Chen, Quansen Sun, Deshen Xia, and David Dagan Feng. Fuzzy c-means clustering with weighted image patch for image segmentation. *Applied soft computing*, 12(6):1659–1667, 2012.

[Jiang *et al.*, 2015] Wenhao Jiang, Feiping Nie, and Heng Huang. Robust dictionary learning with capped l1 norm. In *Twenty-Fourth International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3590–3596, 2015.

[Kannan *et al.*, 2012] SR Kannan, S Ramathilagam, R Devi, and Evor Hines. Strong fuzzy c-means in medical image data analysis. *Journal of Systems and Software*, 85(11):2425–2438, 2012.

[Li and Mukaidono, 1995] Rui-Ping Li and Masao Mukaidono. A maximum-entropy approach to fuzzy clustering. In *Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE Int*, volume 4, pages 2227–2232. IEEE, 1995.

[Li *et al.*, 2008] Mark Junjie Li, Michael K Ng, Yiu-ming Cheung, and Joshua Zhexue Huang. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *Knowledge and Data Engineering, IEEE Transactions on*, 20(11):1519–1534, 2008.

[Miyamoto and Umayahara, 1998] Sadaaki Miyamoto and Kazutaka Umayahara. Fuzzy clustering by quadratic regularization. In *Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*, volume 2, pages 1394–1399. IEEE, 1998.

[Namkoong *et al.*, 2010] Younghwan Namkoong, Gyeongyong Heo, and Young Woon Woo. An extension of possibilistic fuzzy c-means with regularization. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–6. IEEE, 2010.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[Nie *et al.*, 2014a] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *ACM Knowledge Discovery and Data Mining*, pages 977–986, 2014.

[Nie *et al.*, 2014b] Feiping Nie, Jianjun Yuan, and Heng Huang. Optimal mean robust principal component analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1062–1070, 2014.

[Özdemir and Akarun, 2002] Doğan Özdemir and Lale Akarun. A fuzzy algorithm for color quantization of images. *Pattern Recognition*, 35(8):1785–1791, 2002.

[Pham, 2001] Dzung L. Pham. Spatial models for fuzzy clustering. *Computer Vision and Image Understanding*, 84(2):285–297, 2001.

[Ruspini, 1969] Enrique H Ruspini. A new approach to clustering. *Information and control*, 15(1):22–32, 1969.

[Smyth, 2000] Padhraic Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72, 2000.

[Stelios and Vassilios, 2010] Krinidis Stelios and Chatzis Vassilios. A robust fuzzy local information c-means clustering algorithm. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 19(5):1328–1337, 2010.

[Wang *et al.*, 2013] Zhimin Wang, Qing Song, Yeng Chai Soh, and Kang Sim. An adaptive spatial information-theoretic fuzzy clustering algorithm for image segmentation. *Computer Vision and Image Understanding*, 117(10):1412–1420, 2013.

[Yu and Yang, 2007] Jian Yu and Miin-Shen Yang. A generalized fuzzy clustering regularization model with optimality tests and model complexity analysis. *Fuzzy Systems, IEEE Transactions on*, 15(5):904–915, 2007.

[Zhang *et al.*, 2003] Dao-Qiang Zhang, Songcan Chen, Zhi-Song Pan, and Ke-Ren Tan. Kernel-based fuzzy clustering incorporating spatial constraints for image segmentation. In *Machine Learning and Cybernetics, 2003 International Conference on*, volume 4, pages 2189–2192. IEEE, 2003.

[Zhao *et al.*, 2011] Feng Zhao, Licheng Jiao, and Hanqiang Liu. Fuzzy c-means clustering with non local spatial information for noisy image segmentation. *Frontiers of Computer Science in China*, 5(1):45–56, 2011.