International Journal of Approximate Reasoning 106 (2019) 89–106

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Diversity-induced fuzzy clustering

Honghong Cheng, Yuhua Qian *, Yan Wu, Qian Guo, Yong Li

Institute of Big Data Science and Industry, Key Laboratory of Computer Intelligence and China Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi Province, China

ARTICLE INFO

Article history:
Received 6 June 2018
Received in revised form 5 December 2018
Accepted 12 December 2018
Available online 11 January 2019

Keywords:
Granular computing
Diversity
Dependence measure
Fuzzy C-Means

ABSTRACT

Granular computing plays an important role in human reasoning and problem solving, a reasonable granulation method is important in practical tasks. Clustering is one of the most common methods of granulation, learning clear and correct grouping structure of a data set is a key pursuit for clustering algorithm. An excellent clustering algorithm needs to not only explore similar characteristics of individual group but also to pay attention to ensure higher discrimination among different centers. Ignoring the between-cluster variation will lead to a phenomenon that multiple learned centers concentrate to one point, it happens especially when confronted with datasets exist overlapping regions among clusters. To overcome this issue, we model the diversity information in-between different clusters and measure it with a statistical dependence metric Hilbert Schmidt Independence Criterion (HSIC), and then develop a Diversity-induced Fuzzy C-Means clustering algorithm framework based on traditional Fuzzy C-Means algorithm, which can minimize the within-cluster dispersion and maximize between-clusters separation simultaneously. The formula of updating center attracts the points have the same group with it as well as excludes the impact from other clusters. We analyze the convergence of proposed method under the alternating minimizing optimization fashion, and discuss the sensitivity of parameters in algorithm for clustering performance. The reasonability and advantages of proposed method also have been explained by simulation study. Further, three types of DiFCM methods by using different HSIC form carry out on UCI and image data sets, all experimental results confirm the outstanding of the proposed method.

1. Introduction

Granular computing (GrC) plays a fundamental role in human reasoning and problem solving which originally proposed by Zadeh [1]. The idea has been applied in many fields such as machine learning, databases, data mining and knowledge discovery [2–5]. Clustering analysis is one of the popular methods of granulation, often used to discovery the inherent grouping structure in a set of subjects and has extensive applications in many different areas such as image processing [6], information retrieval [7], multi-modal data analysis [8], scientific data exploration [9]. The goal of a clustering algorithm is to group a set of unlabeled objects into several meaningful clusters so that the objects in the same cluster have relatively high similarity and in different clusters have very high dissimilarity. In this respect, clustering can give us an insight into

* Corresponding author.
E-mail addresses: chhsxdx@163.com (H. Cheng), jinchengqyh@126.com (Y. Qian), 18335103184@163.com (Y. Wu), czguoqian@163.com (Q. Guo), ly.mrty@gmail.com (Y. Li).

https://doi.org/10.1016/j.ijar.2018.12.010
0888-613X/

the distribution of the data and show the clear group structures. Over the past few decades, a host of clustering algorithms have been developed for different clustering tasks [10–13].

Many of those clustering methods, between-cluster information which contributes to obtain a distinct between-cluster separation is often neglected, hence leads to a phenomenon that few clustering centers are represented the same or similar clusters. In this article, we try to overcome the issue to some extent. We realize our idea based on fuzzy C-Means algorithm (FCM), which is introduced by Bezdek [14] in 1981 as a distinguished and representative soft clustering method [15–17]. FCM allows each point to have memberships in all clusters rather than hard C-Means just assigns a point to one distinct cluster. So the membership matrix provides more information in the uncertain way than it does in the certain case [18–21].

In FCM algorithm, the maximal membership of an observation determines the cluster that the point belongs to, but the computing formula of membership only depends on the distance between the observation and c cluster centers. If cluster centers are not accurate enough, they can affect the membership values, the maximal membership of single sample will choose an improper cluster in turn. In fact, the cluster center is not only effected by the shape, size and distribution of the individual group, but also the influenced by the distribution of other clusters. Ignoring the variation between different centers will lead to a phenomenon that multiple learned centers concentrate to one point, it happens especially when confronted with datasets exist overlapping regions among clusters. In order to overcome the shortage, many studies have been developed from different views [13,22–25]. C.J. Veenman et al. imposed a hard constraint on the cluster variance based on the hypothesize that clusters will cooperate with neighboring clusters to make the demarcation of clusters more distinct [13]; Liang Bai et al. proposed to directly penalize the pairwise similarities between categorical cluster centers in the task of categorical data clustering [23]; Weiling Cai et al. incorporated local spatial on the membership functions to reduce the noise and outliers for improving the image segmentation precision [22]; Liyong Zhang et al. reconstructed the supervised information from the original data and introduced dual expression between cluster prototypes to FCM clustering, provided a reconstructed data Fuzzy C-means clustering to improve the FCM clustering performance [25]. Most past efforts have been spent on only augmenting intuitive motivation of clustering task with minimized within-cluster distances and maximized between-clusters separations, and less considering the representativeness of cluster centers. Thereupon, we will fetch the diversity information among clusters through their centers.

Some researchers also have noticed this view when studying image clustering, they state that the “cluster one-sidedness” problem which leads some algorithm fail to identify the adjacent small clusters, and in order to overcome the issue, they imported the angle between centers to enhance the diversity of different image centers [26]; Heiko Timm et al. forced centers away from each other by introducing a mutual repulsion concept between centers to avoid the drawback [20], though their research based on the possibility fuzzy clustering where the membership can be 0; based on the advantage of possibility, adaptive PCM is proposed to adapt the cluster number by introducing uncertainty into membership function [27]. These all reflect the importance of diversity information during clustering from different perspectives.

Inspired by the effectiveness of these algorithms, in this paper, we fetch between-clusters diversity information via making different cluster centers representative as much as possible, so that they can possess more own cluster information as well as eliminate other cluster impact. We bring in a statistical independent index Hilbert Schmidt Independence Criterion (HSIC) to enforce the diversity between cluster centers [28], and develop a series of diversity induced Fuzzy C-Means clustering method, called Diversity-induced Fuzzy C-Means algorithm (we will rewrite it as DiFCM for the sake of convenience in below sections). DiFCM combines the original Fuzzy C-Means paradigm and diversity regularization term of in-between cluster centers. With the diversity regularization term, we explore the diversity of different clusters especially in the case that the distinction between one cluster center and other center is weak. Because HSIC is constructed in the Reproducing Kernel Hilbert Space (RKHS), we can choose a lot of kernel functions to compute the HSIC, here we select three common kernel functions to mining the diversity information. The detailed analysis of the diversity regularization term will be exhibited in section 3. The major contributions are as follows:

- We first introduce the statistical dependence index HSIC to measure the diversity information of different clusters by considering the representativeness of a individual cluster center and the distinctiveness between different cluster centers.
- The three updating formulas of the cluster center are obtained and unique closed solutions are offered for proposed DiFCM with three different forms of diversity regularization terms.
- The three types of proposed methods are compared with four baselines over six UCI data sets and six image data sets, the experimental results favorably outperform other methods. As for our methods, the DiFCM with diversity regularization terms constructed by linear kernel is superior to other two kinds.

The rest of this paper is organized as follows. The Fuzzy C-Means algorithm is briefly analyzed in section 2. In section 3, we introduce the HSIC to enforce the between clusters separation by control diversity information between cluster centers. In section 4, the DiFCM with diversity regularization term formed by linear kernel function as a special case is analyzed and solved by alternating minimizing optimization fashion and updating formulas of two methods with polynomial and sigmoid kernel functions are also displayed. In section 5, several experiments are illustrated to perform the proposed algorithm. Finally, a conclusion is reached in section 6.

2. Fuzzy C-Means clustering algorithm

Assuming that $X = \{x_1, x_2, \cdots, x_n\} \in R^{p\times n}$ is the matrix of data set, where each column is a sample and p is the dimensionality of the feature space, that is $x_i = (x_{i1}, \cdots, x_{ip})', i = 1, \cdots, n$. The fuzzy C-Means algorithm partitions a data set X into c clusters which can be characterized by centers matrix $C = \{c_1, c_2, \cdots, c_c\} \in R^{p\times c}$ and $c_i = (c_{i1}, \cdots, c_{ip})', i = 1, \cdots, c$ is the cluster center of *i-th* corresponding cluster. With the help of membership u_{ij}, each observation x_i obtains the probability belonging to all clusters c_j, the cluster centers are updated by the sum of normalized weighed samples. Through the process, the homogeneous elements are divided into distinct subsets [29].

In the FCM algorithm, the objective is to search the U and C that minimize the summation of weighted distances:

$$f_{FCM}(X; U, C) = \sum_{i=1}^{n}\sum_{j=1}^{c} \mu_{ij}^{m}\|x_i - c_j\|^2 \tag{1}$$

with the constrains:

$$\begin{cases} u_{ij} \in (0, 1], \quad 1 \le i \le n, 1 \le j \le c, \\ \sum\limits_{j=1}^{c} \mu_{ij} = 1, \quad i = 1, \cdots, n, \\ 0 < \sum\limits_{i=1}^{n} \mu_{ij} < n, \quad j = 1, \cdots, c, \end{cases} \tag{2}$$

where c meets $2 \le c \le n$, the parameter $m \in (1, \infty)$ is the fuzzy index which influences the fuzziness of the partition. $U = [u_{ij}]$ is a $n \times c$ real matrix, u_{ij} is the membership of x_i belongs to the c_j. $\|\cdot\|^2$ is the squared Euclidean distance, we will denote that $d^2(x_i, c_j) = \|x_i - c_j\|^2 = \sum\limits_{k=1}^{p}(x_{ik} - c_{jk})^2$ in the following sections.

The objective function with the constraints (2) is a constrained nonlinear optimization problem, which usually can be solved by alternating optimization strategy [30]. In this optimization framework, one usually fixes one parameter to find the optimal solution of the other parameters satisfying the constraints [30,31]. Specially, the processing can be considered as solving two subproblems:

Subproblem 1. Fix C to be constant, the membership matrix update equation $f(U, \hat{C})$ can be solved by

$$\frac{\partial f^L(U, \hat{C})}{\partial U} = 0 \tag{3}$$

Subproblem 2. Fix U to be constant, the cluster center update equation $f(\hat{U}, C)$ can be solved by

$$\frac{\partial f^L(\hat{U}, C)}{\partial C} = 0 \tag{4}$$

where f^L represents the objective function (1) processed by Lagrange multipliers.

According to the above analysis, the membership can be obtained as follows:

$$u_{ij} = \frac{1}{\sum\limits_{k=1}^{c}\left(\frac{d^2(x_i, c_j)}{d^2(x_i, c_k)}\right)^{\frac{1}{m-1}}} \tag{5}$$

the cluster center is updated with:

$$c_j = \frac{\sum\limits_{i=1}^{n} u_{ij}^{m} x_i}{\sum\limits_{i=1}^{n} u_{ij}^{m}} \tag{6}$$

The Fuzzy C-Means clustering algorithm move the cluster centers to the right location within a given data set in an iterative fashion. The real cluster that an observation belongs to completely depends on the maximal value of its membership. But they only consider the within-cluster information and ignore the information in-between clusters, this will lead to a poor separation especially in the case that the distinction between one cluster and other clusters is weak induced by big overlapping region among clusters, however, one still hopes to obtain a clear partition for those points lie in overlap region. In the centroid-based clustering algorithms, the cluster information can be well represented by its cluster center, so we

can control the between-clusters information by controlling the relationship between clusters. On the other hand, maximal membership $u_{ij}, j = 1, \cdots, c$ decides a variate x_i to a most likely cluster c_j and the computing of membership depends on all the cluster centers, however, the updating form of centers in formula (6) only considers the information within clusters. Modifying the updating form of centers may improve the judgment of membership indirectly.

3. Diversity-induced Fuzzy C-Means algorithm

According to the above analysis, it's meaningful to focus on updating of cluster centers. In the paper, we seek to minimize dependence between different clusters, small dependence of two cluster centers means high diversity between them, we deem the diversity information of a cluster is a cluster own some special information different from other clusters. We employ an independence criterion Hilbert–Schmidt Independence Criterion (HSIC) to measure the independence between cluster centers, which has been applied to clustering problems many times [32–34], for some advantages. First, it guarantees good uniform convergence theoretically and has little bias even in high dimensions. Second, it can capture much complex nonlinear dependence and needn't to explicitly consider the joint distribution of random variables. Third, its empirical estimation is the trace of product of cluster centers, which makes objective function easy to solve. In this article, to ensure that the cluster centers in different clusters provide enough diversity information respectively, we introduce it to penalize the dependence between two cluster centers.

In order to build the diversity regularization term among cluster centers, we first review the definition of HSIC [28].

3.1. Measure of diversity

Assume that $\mathcal{F}$ is the Reproducing Kernel Hilbert Space (RKHS) on $\mathcal{X}$ with associated kernel $k : \mathcal{X} \times \mathcal{X} \to R$, and feature map $\phi : \mathcal{X} \to \mathcal{F}$. That is, for each point $x \in \mathcal{X}$, there is a corresponding element $\phi(x) \in \mathcal{F}$ such that $\langle \phi(x), \phi(x') \rangle_{\mathcal{F}} = k(x, x')$, where $k : \mathcal{X} \times \mathcal{X} \to R$ is a unique positive definite kernel. Let $\mathcal{G}$ be the RKHS on $\mathcal{Y}$ with kernel l and feature map ψ, and $\langle \psi(y), \psi(y') \rangle_{\mathcal{G}} = l(y, y')$ also holds. Then, we will have the definition of cross-covariance operator $C_{xy} : \mathcal{G} \to \mathcal{F}$.

$$C_{xy} = E_{xy}[(\phi(x) - u_x) \otimes (\psi(y) - u_y)] \tag{7}$$

where $u_x = E[\phi(x)]$, $u_y = E[\psi(y)]$, and $\otimes$ is the tensor product. The definition of HSIC is as follows:

Definition 1. Given two separable (RKHSs) $\mathcal{F}, \mathcal{G}$ and a joint distribution measure p_{xy} over $(\mathcal{X}, \mathcal{Y})$, we define the Hilbert–Schmidt Independence Criterion (HSIC) as the squared Hilbert–Schmidt norm of the associated cross-covariance operator C_{xy}:

$$HSIC(p_{xy}, \mathcal{F}, \mathcal{G}) = \|C_{xy}\|^2_{HS} \tag{8}$$

where $\|\cdot\|_{HS}$ denotes the Hilbert–Schmidt norm of a matrix.

Proposition 1. *Given two random variables $X \sim p$ and $Y \sim q$, with joint distributions p_{XY}, and two RKHS's $\mathcal{F}$ and $\mathcal{G}$ with characteristic kernels k and l, then $HSIC(p_{XY}, \mathcal{F}, \mathcal{G}) = 0$ if and only if $p_{XY} = pq$, i.e. if X and Y are independent [35].*

The Definition 1 and Proposition 1 indicate that HSIC is non-negative index, and the smaller the value is, the more independent between two variables is.

Now we adopt an easy estimated formulation of HSIC. Given a series of n independence observations drawn from p_{xy}, $S := \{(x_1, y_1), \cdots, (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$, we obtain an empirical version of HSIC, rewritten as $HSIC(S, \mathcal{F}, \mathcal{G})$, is given by

$$HSIC(S, \mathcal{F}, \mathcal{G}) = (n-1)^{-2} tr(HKHL) \tag{9}$$

where $K, L \in R^{n \times n}$ are the Gram matrices, with $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$. H is a centralizing matrix which centralizes the Gram matrix to have zero column mean in the Hilbert feature space, where $H_{ij} = \delta_{ij} - \frac{1}{n}$ and $\delta_{ij} = 1$ if only $i = j$, otherwise $\delta_{ij} = 0$, tr(X) is trace operation in linear algebra which sums the diagonal elements of matrix X. For more details of HSIC, one can refer to the paper [28,35]. The larger the HSIC value is, the closer the dependence relationship between two variables is. Leveraging this property, in our paper, we treat different cluster centers as different variables, minimize HSIC value between two centers as much as possible to separate two cluster centers, namely, the diversity between two clusters is maximized.

3.2. The proposed methods

To enhance the diversity information in different clusters, we encourage the centers of different clusters to be of sufficient diversity with each other. This amounts to enforcing the data distributes of different clusters to be novel with each other. HSIC is used to measure the diversity of two cluster centers, so the diversity regularization term contains $c(c-1)$ diversity

Table 1
Kernel functions.

Kernel function	Form	$HSIC(c_i, c_j)$ in this paper
Linear	$k(x, y) = x^T y + c$	$tr(Hc_i c_i^T Hc_j c_j^T)$
Polynomial	$k(x, y) = (\rho(x^T y) + c)^d$	$tr(H(c_i c_i^T + 1)^d H(c_j c_j^T + 1)^d)$
Sigmoid	$k(x, y) = tanh(\ \rho(x^T y) + c)$	$tr(Htanh(\rho(c_i c_i^T) + 1)Htanh(\rho(c_j c_j^T) + 1))$

Table 2
Updating formulas in DiFCM algorithm.

λ	DiFCM		$M_j =$	$K_l =, l = 1, \cdots, c$
	Methods	$c_j =$		
$\lambda = 0$	FCM	$\frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}$		
$\lambda \neq 0$	DiFCM-lin	$(\sum_{i=1}^{n} \mu_{ij}^m I + \lambda \frac{1}{(p-1)^2} M_j)^{-1} (\sum_{i=1}^{n} \mu_{ij}^m x_i)$	$\sum_{i \neq j}^{c} HK_i H$	$c_l c_l^T$
	DiFCM-pol	$(\sum_{i=1}^{n} \mu_{ij}^m I + d\lambda \frac{1}{(p-1)^2} M_j K_j^{\frac{d-1}{d}})^{-1} (\sum_{i=1}^{n} \mu_{ij}^m x_i)$		$(c_l c_l^T + 1)^d$
	DiFCM-sig	$(\sum_{i=1}^{n} \mu_{ij}^m I + \rho\lambda \frac{1}{(p-1)^2} M_j (1 - K_j^2))^{-1} (\sum_{i=1}^{n} \mu_{ij}^m x_i)$		$tanh(\rho(c_l c_l^T) + 1)$

pairs for all c cluster centers. Importing the diversity regularization term to the original Fuzzy C-Means clustering objective function, we will minimize the following objective function:

$$f_{DiFCM}(X; U, C) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^m \|x_i - c_j\|^2 + \lambda \sum_{i=1}^{c} \sum_{j \neq i} HSIC(c_i, c_j) \tag{10}$$

$$\text{s.t.} \sum_{j=1}^{c} \mu_{ij} = 1, \mu_{ij} > 0, i = 1, \cdots, n, \tag{11}$$

where $\lambda \geq 0$ is a tradeoff parameter which balances the within-cluster similarity and between-cluster diversity. When $\lambda = 0$, the diversity term plays no role in the clustering process.

In the objective function (10), the first item ensures the minimal distance loss of the same cluster, the second item guarantees cluster centers meet maximal diversity with others. According to the formula (10), we can choose a kernel freely which allows us to incorporate prior knowledge into the dependence estimation process. In this paper, three typical kernel functions are adopted to study the diversity information, the details are listed in the Table 1. For convenience, if we use linear kernel function to measure the HSIC, the corresponding objective function is denoted by DiFCM-lin, the other two are similar to linear, denoted by DiFCM-pol and DiFCM-sig respectively and we provide the updating formula of U and V for three objective functions in the Table 2.

From the Table 1, we can find that three kernel functions have a common inner element $x^T y$, the difference among them is their outer nonlinear functions after the inner product of two variables, which is a critical step for those kernel functions. Hence, we analyze the DiFCM-lin as a special case:

$$f_{DiFCM-lin}(X; U, C) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^m \|x_i - c_j\|^2 + \lambda \frac{1}{(p-1)^2} \sum_{i=1}^{c} \sum_{j \neq i} tr(Hc_i c_i^T Hc_j c_j^T) \tag{12}$$

$$\text{s.t.} \sum_{j=1}^{c} \mu_{ij} = 1, \mu_{ij} > 0, i = 1, \cdots, n. \tag{13}$$

Analyzing the second term of formula (12), according to the exchangeability quality of trace operation [36] and the property of idempotent matrix [37], we have $tr(Hc_i c_i^T Hc_j c_j^T) = tr(c_i^T Hc_j c_j^T Hc_i) = (c_i^T Hc_j)^2 = (c_i^T H^T Hc_j)^2 = ((Hc_i)^T Hc_j)^2$, where H is an idempotent matrix with the property $H * H = H$. It seems that diversity term is the sum square of cosine

similarity measure between two cluster centers, but we have two advantages over the cosine similarity. First, every cluster center has been scaled by centralized matrix H so that compared cluster centers have 0 mean, this means that we obtain the variation of similarity just induced by fluctuations of every cluster center, it will be more precise than untreated cluster centers. Second, cosine similarity will have positive and negative value, but we need to penalize on both high positive and negative dependence, using the absolute value of cosine similarity is one of choice, however, it will bring some troubles during optimizing the proposed objective, so we control the diversity information between cluster centers have the same form with squared cosine similarity. The effectiveness of positive constrain has been verified in [26] where the acute angle between each pair of centers bas been considered. The smaller value means two cluster centers prefer to be more independent, in other words, each of two cluster centers try to own its unique information, this will guide the boundary points to determine a more differentiated cluster.

From the above analysis, we learn the proposed fuzzy clustering methods is reasonable, next, we will offer the optimal solution for DiFCM-lin method.

4. Solving and analyzing the proposed objective function

In this section, we will make efforts to solve the objective function in equation (12) which is not convex in both U and C [30]. Therefore, it's unrealistic to expect an algorithm to guarantee the global minimum solution [30,38]. The function carries four variables: fuzzy index m, tradeoff parameter λ, membership matrix U and cluster centers matrix C. We will regard m and λ, which depend on the data set, as constants in the process of solving problem. Then we only deal with two variables U and C, it's natural to apply an alternating minimizing strategy with that optimizing the function with respect to one variable while fixing the other one [23,31]. So the complex function will be reduced to two manageable subproblems.

Subproblem 3. As one will see, the second part in equation (12) has nothing to do with the U. So fix C to be a constant, the updating equation of membership is the same as the formula (5) in Subproblem 1. Specifically, the objective function is a constrained linear optimization problem with respect to each u_{ij}, which can be solved by Lagrange Multipliers Method easily.

$$\frac{\partial f^L_{DiFCM-lin}(U,\hat{C})}{\partial u_{ij}} = \frac{\partial f^L(U,\hat{C})}{\partial u_{ij}} = 0. \tag{14}$$

Subproblem 4. Fix U to be a constant, we need to solve c cluster centers in problem (12), they all have the same status. Without losing generality, the computing details of j-th cluster center will be chosen to show as an example. The sub-objective function about c_j is as follows:

$$\begin{aligned} f^L_{DiFCM-lin}(X;\hat{U},c_j) = &\sum_{i=1}^{n}\mu_{ij}^m\|x_i-c_j\|^2 \\ &+\lambda\frac{1}{(p-1)^2}\sum_{i\neq j}^{c} tr(Hc_jc_j^THc_ic_i^T) \end{aligned} \tag{15}$$

According to the property of trace operator, the second term can be represented as $\sum_{i\neq j}^{c} tr(Hc_jc_j^THc_ic_i^T) = tr(c_j^T(\sum_{i\neq j}^{c} Hc_ic_i^TH)c_j)$. Let $\sum_{i\neq j}^{c} Hc_ic_i^TH = M_j$, the equation (15) becomes

$$\begin{aligned} f^L_{DiFCM-lin}(X;\hat{U},c_j) = &\sum_{i=1}^{n}\mu_{ij}^m\|x_i-c_j\|^2 \\ &+\lambda\frac{1}{(p-1)^2} tr(c_j^TM_jc_j) \end{aligned} \tag{16}$$

where $f^L_{DiFCM-lin}$ represents the unconstrained problem corresponding to the transformed $f_{DiFCM-lin}$ by Lagrange Multipliers Method. Problem (16) is a smooth convex program. Differentiating the $f_{DiFCM-lin}(X;U,c_j)$ with respect to c_j and setting it to zero, we will get the following optimal solution c_j that satisfies

$$(\sum_{i=1}^{n}\mu_{ij}^m I + \lambda\frac{1}{(p-1)^2}M_j)c_j = \sum_{i=1}^{n}\mu_{ij}^m x_i \tag{17}$$

where I is an identity matrix and its size is $p \times p$. The equation (17) is a standard linear equation who has a unique solution if the $(\sum_{i=1}^{n} \mu_{ij}^{m} I + \lambda \frac{1}{(p-1)^2} M_j)$ matrix is full rank. The requirement of full rank will hold in generally hereof, if the reverse case happens, we get the inverse of the matrix via a transformation method in footnote.[1] Therefor the cluster center c_j is updated by

$$c_j = (\sum_{i=1}^{n} \mu_{ij}^{m} I + \lambda \frac{1}{(p-1)^2} M_j)^{-1} (\sum_{i=1}^{n} \mu_{ij}^{m} x_i) \tag{18}$$

Compared the computing form of cluster center in formula (18) and formula (6), we can observe that the former not only contains the same information with the latter, but also removes the effect of other cluster centers, it means the current center just keeps the information of corresponding cluster. The implication of the formula (18) is similar to the formula (10) in [20], where a clustering center attracted by the data assigned to it and repelled by the other clusters. The difference between them is measurement approaches of diversity information, our methods encourage large independence between two clustering centers, the method in [20] hopes large distance between two centers.

Moreover, in order to update the cluster centers at each iteration, we choose to initialize membership matrix firstly and use the cluster centers solved by original FCM as initialization centers. The whole detailed procedure of DiFCM-lin is represented in Algorithm 1.

Algorithm 1 Diversity-induced Fuzzy C-Means Algorithm (DiFCM-lin)

Input: A data set X, the number of data clusters c, difference error δ, maximal number of iteration T, fuzzy index m, tradeoff parameter λ.
Output: Converged U and C
Step1: Randomly initialize membership matrix U where u_{ij} meets the formula (2), then obtain the initialization cluster centers $C^1 = \{c_1^1, \cdots, c_c^1\}$ by Algorithm 1.
Step2: Computing $U^1 = [u_{ij}^1]$ with formula (16) such that $f_{DiFCM-lin}(X; U, \hat{c}_j)$ is minimized, let $t = 1$.
Step3: Update the cluster centers c_j^{t+1} with formula (5) such that $f_{DiFCM-lin}(X; \hat{U}, c_j)$ is minimized one by one, then reach the $f_{DiFCM-lin}(X; \hat{U}, C)$ minimum.
If $\|f_{DiFCM-lin}(U^t, C^{t+1}) - f_{DiFCM-lin}(U^t, C^t)\| \leq \delta$ or $t > T$, then stop; otherwise go to Step4.
Step4: Update membership matrix U^{t+1} with formula (18) such that $f_{DiFCM-lin}(U^{t+1}, C^{t+1})$ is minimized.
If $\|f_{DiFCM-lin}(U^{t+1}, C^{t+1}) - f_{DiFCM-lin}(U^t, C^{t+1})\| \leq \delta$ or $t > T$, then stop; otherwise, $t = t + 1$, go to Step3.

The Algorithm 1 is carried out until convergence with a finite number of iterations. For the DiFCM-pol and DiFCM-sig methods, we only just set the corresponding clustering center updating formulas from Table 2 into the Algorithm 1.

Next, we will discuss the time complexity and convergence of proposed methods theoretically.

4.1. Time complexity analysis

The proposed fuzzy clustering algorithm is scalable to the number of objects, dimensions and clusters. We only consider the two major computation processes. The computation cost for U is the same as the original FCM which is $O(npc)$ [23]. As for the computational complexity for updating cluster centers matrix C, it should include two parts, one is the computing cost of M which is $O(p^2 c)$, the other part is inverse operation in formula (18), here the inverse of matrix is obtained by definition, to which, the computing cost is $O(p^3 c)$, so the time complexity is $O(p^3 c)$. If the iterative times is t when the algorithm stops, the total computational cost of proposed method is $O(npct + p^3 ct)$. It shows that computational complexity is linear for the number of objects and clusters, and nonlinear for the number of dimensions. It is acceptable on many data sets because of its better clustering performance and $n > p$ in many cases.

4.2. Algorithm convergence analysis

Algorithm 1 has the same optimization framework as original FCM, so it is easy to prove the convergence.

Theorem. *The objective function (12) is guaranteed to convergence with Algorithm 1.*

Proof. The objective function has been divided into two subproblems.

1) The optimization procedure of Subproblem 3 is just the same as the Subproblem 1. With the Lagrange multipliers technique, U obtained by the formula (14) is the optimal minimum solution for Subproblem 1, which has been demonstrated in many literatures [14,39–42], we will not go into in this paper.

[1] The matrix can be simplified as $A + CBC'$, where $A = \sum_{i=1}^{n} \mu_{ij}^{m} I$, $B = \lambda \frac{1}{(p-1)^2} I$, $C = \sum_{i \neq j}^{c} Hc_i$, then $(A + CBC')^{-1} = A^{-1} - A^{-1}C(B^{-1} + C'A^{-1}C)^{-1}C'A^{-1}$, if $p \neq 1$, both A and B exist the inverse matrix, so we can obtain the inverse matrix of it.

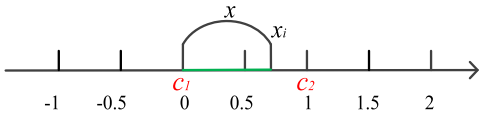


Fig. 1. The legend of distance between x_i and c_1.

2) For the Subproblem 4, it can be considered as c sub-objective functions (15). From the analysis for the formula (15), we can obtain the unique closed solution as equation (17). The cluster centers matrix C is composed of c_j, $j = 1, \cdots, c$ by column gradually, so the Subproblem 4 also has a unique closed solution.

Combining two parts 1) and 2), it forms an iterative optimization method to minimize the objective function (12) at each iteration. Then, according to the alternating minimizing strategy, the following inequality sequence will be set up:

$$\begin{aligned} f_{DiFCM-lin}(U^{t-1}, C^{t-1}) &\geq f_{DiFCM-lin}(U^{t}, C^{t-1}) \\ &\geq f_{DiFCM-lin}(U^{t}, C^{t}) \geq f_{DiFCM-lin}(U^{t}, C^{t+1}) \end{aligned} \tag{19}$$

Therefore, for each iteration, the objective function is non-increasing. Hence, the convergence of DiFCM is proved.

Remark. One will notice that the solution yielded via alternating optimization is not a globe optimization, however, the solutions of Subproblem 3 and Subproblem 4 are closed uniquely, therefore the solution to Equation (12) is local but unique.

As for the DiFCM-pol and DiFCM-sig methods, we can get the same property of convergence under the same iterative optimization strategy.

5. Experimental results

In this section, we highlight the properties of the DiFCM-lin method on a 3D synthetic data and a modified Iris data. And then several baseline clustering algorithms are compared with three types of proposed methods to demonstrate the validity and feasibility of the diversity information on twelve real datasets. Furthermore, we demonstrate the convergence performance and sensitivity of parameters in DiFCM-lin method.

5.1. Baseline algorithm and parameters setting

We compare following four clustering algorithms to illustrate the superiority of DiFCM.

- FCM [14]: Fuzzy C-Means is a popular fuzzy clustering algorithm, which is the main algorithm that needs improvement and comparison in the article.
- KFCM [43]: Kernel fuzzy C-Means transfers the original data space into kernel space and analyzes clustering performance on the kerneled features. In this paper, Gaussian kernel function is used to compare with our method.
- PoFCM [20]: Modified possibilistic fuzzy c-means clustering is based on the framework of possibilistic clustering, which is different from the traditional probabilistic fuzzy C-Means, where the constrain that the sum of membership degrees for each data equals one is dropped.
- DrFCM [26]: Diverse fuzzy C-Means introduces a diversity regularization into the traditional fuzzy C-Means in order to encourage the cluster centers to cover more information in data space.

All of the above methods are based on the traditional fuzzy C-Means, the selection of fuzzy index m is their common problem. Rigorous mathematical arguments for learning and determining a suitable one with respect to a specific data set is very difficult, and it still remains unknown in literatures so far. Herein, we discuss the choice of m in the light of some heuristic guidelines in some articles [44–47], and give a reasonable interpretation for the setting of m in this paper by demonstrating in Example.

Example. Assuming that there are only two clusters, c_1 and c_2 denote cluster centers respectively, the distance between them is 1. Let the distance between sample x_i and cluster center c_1 is x, which is displayed in Fig. 1, the membership is $u_{i1} = \frac{1}{|\frac{x}{1-x}|^{\frac{2}{m-1}} + |\frac{x}{x}|^{\frac{2}{m-1}}} = \frac{1}{|\frac{x}{1-x}|^{\frac{2}{m-1}} + 1}$, we plot a bunch of curves versus the fuzzy index m ranging from 1.1 to 4.1 and the step length is 0.2 in Fig. 2.

Fig. 2 depicts that the bigger the m is, the smaller the u_{i1} is for a fixed $x \in [-1, 0.5]$. Conversely, the smaller the m is, the smaller the u_{i1} is for a fixed $x \in [0.5, 2]$. The curve drew by $m = 2.9$ is a watershed for all curves with respect to curve rate: bold red ($m = 2.9$), bold green ($m = 3.1$) and bold blue ($m = 2.7$) curves have shown the variation. The slope of curves by $m > 2.9$ is steeper than the slope of curves by $m < 2.9$ when $x \to 0.5^-$ and $x \to 0.5^+$, it means that the u_{i1} obtained by $m > 2.9$ approaches 0.5 more quickly than the u_{i1} obtained by $m < 2.9$ when x is close to 0.5. However, it is ambiguous to

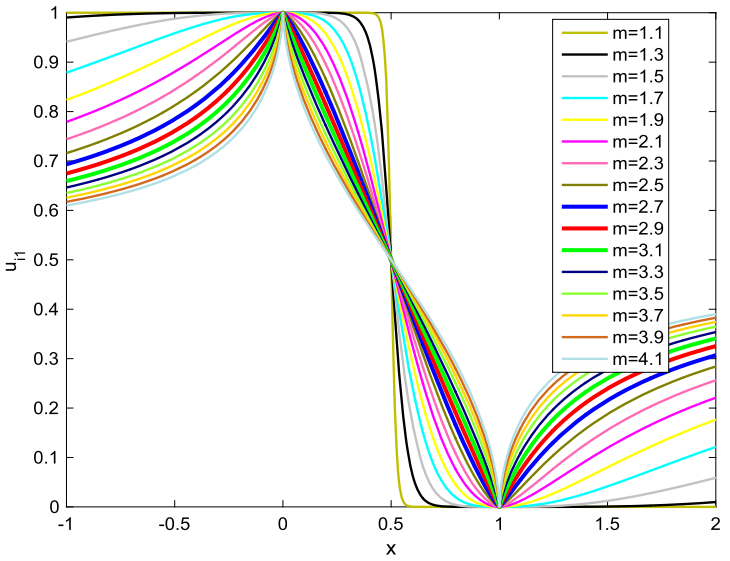


Fig. 2. The membership u_{ij} with varying m values. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

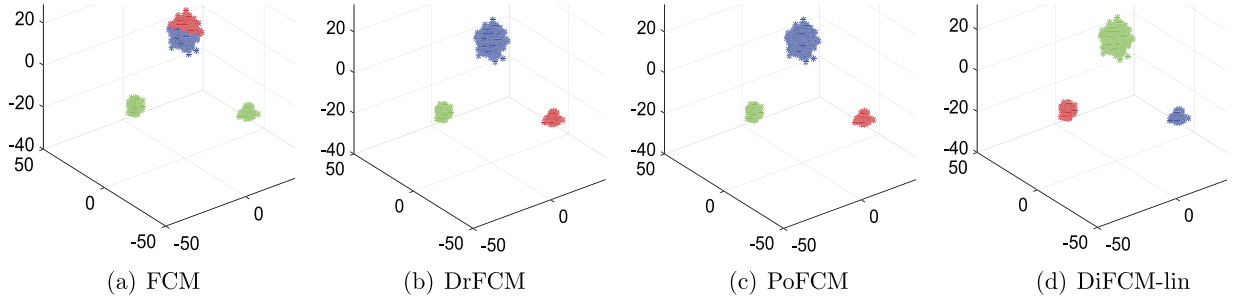


Fig. 3. Demonstration of baselines and DiFCM-lin on 3D synthetic data.

make a clear data partition when $u_{i1} = 0.5$, we should try to prevent it happening. Moreover, the data point x_i completely belongs to c_1 (c_2) when $x \in [-1, 0]$ ($x \in [1, 2]$), the u_{i1} value should be larger than 0.5. Based on above analysis, we traverse $m \in [1.1, 2.9]$ with step-size 0.2 in our experiments to find optimal solution.

Except for common fuzzy index parameter, each of compared method has their own parameters, those parameters seriously depend on special data sets, grid searching within an appropriate range of related parameters is a common and effective method [26,27,43]. So for KFCM, the standard deviation in Gaussian kernel function is assigned in $[0.1, 1]$ and $[10, 100]$ two intervals. For PoFCM, which includes two parameters: weighting factor γ and tolerable minimum distance between two neighboring clusters η, we set the γ vary from $\{10^{-2}, \cdots, 10^{2}\}$, the η vary from $\{10^{-3}, \cdots, 10^{1}\}$ following the original settings. In DrFCM, regularization parameter λ and step size ρ, are tuned in $\{10^{-5}, \cdots, 10^{-1}, 10^{0}\}$ referring to the original literature. As for our method, the regularization term λ is searched in sequence $\{10^{-5}, \cdots, 10^{-1}, 10^{0}, 10^{1}, \cdots 10^{5}\}$. The properties of kernel functions and related parameter settings for machine learning and pattern analysis have been discussed in some literatures [48,49], in this paper, for DiFCM-pol the exponent d in polynomial kernel function is assigned in $\{2, 3\}$ and for DiFCM-sig, the scale factor ρ is set vary from 1 to 10 with step-size 1 heuristically.

5.2. Highlight the properties

5.2.1. Demonstration on synthetic data

We compare the traditional FCM and two modified FCM baselines with our proposed method DiFCM-lin to illustrate its effectiveness on a tested 3D synthetic data [26]. It contains three clusters, the size of them is very different, the biggest group contains 3000 samples, the other groups contain 200 and 100 separately. The clustering performances on unbalanced datasets like this are usually influenced if the clustering methods ignore the scale difference among clusters.

Not surprisingly, the traditional FCM trends to divide the large cluster into two groups, and combines two originally small clusters into one group in Fig. 3(a). The blue, red and green stars respectively represent the three learned clusters in Fig. 3(b), (c), (d), we find that all modified FCM algorithms including ours can perfectly identify the true group, which owns the consideration of the diversity between different cluster centers.

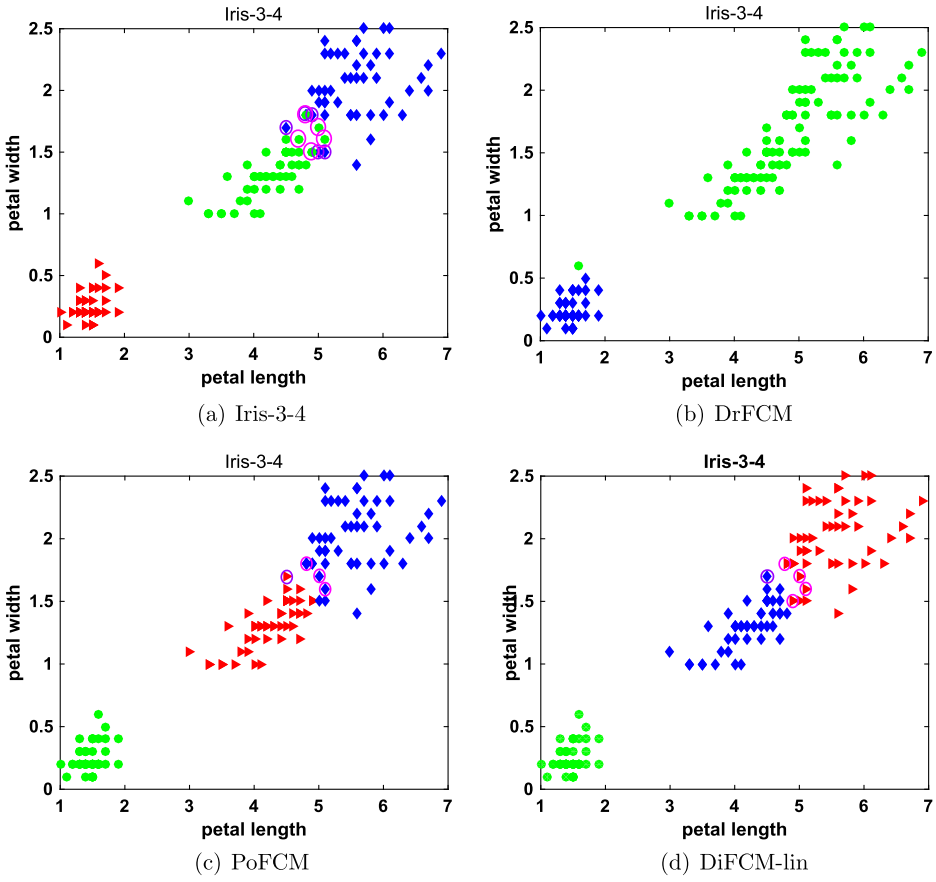


Fig. 4. Demonstration of baselines and DiFCM-lin on Iris-3-4 data.

5.2.2. Demonstration on modified Iris data

In this section, we exhibit clustering performances on data set where exists some overlap among clusters. We use the well-known Iris data only with the petal length and the petal width as in [20], for convenience, we note it as Iris-3-4. Fig. 4(a) shows the distribution information of Iris-3-4, the red triangles, the green hexagons and the blue diamonds represent three true clusters, the five blue points circled by purple circle and the five green points circled by pink circle are located in the overlap region of the two clusters. We conduct three methods including our method on Iris-3-4 and show clustering performance in Fig. 4(b), (c) and (d), where the red, green and blue geometries are the learned clusters. The subfigure (b) displays the results of DrFCM method, it seems only two clusters are learned, because the green cluster merges the two adjacent groups. The possible reason for this result is DrFCM pays too much emphasis on the similarity of cluster shapes during the clustering process. In contrast, the results of PoFCM and DiFCM-lin perform better, they can more correctly separate the points in overlap region to right groups. The points circled in subfigure (c) and (d) don't get their correct groups, these points are not more than half of the original points in confusion area.

5.3. Demonstration on real data

5.3.1. Data sets description

The above subsection quantifies the originality of our methods by simulation study. Next, we carry out experiments on real data sets to show advantages and practicability of suggested method. Six data sets downloaded from UCI Machine Learning Repository [50] and six benchmark image data sets are conducted in our experiment phase [26,51]. The UCI data sets are outlined in Table 3, and image datasets are detailed as follows:

- JAFFE: The Japanese Female Facial Expression database contains seven facial expressions: happy, angry, disgust, fear, sad, surprise and neutral, those are posed by 10 Japanese females. Each face image is represented by 676-dimensional gray pixel values. The scale of it is 213 Objects, 676 Dimensions and 10 Classes.
- HandWritten: HandWritten dataset contains 0 to 9 ten handwritten digits, each image is represented by 240-dimensional gray pixel values. The scale of it is 2000 Objects, 240 Dimensions and 10 Classes.

Table 3
Six data sets in the experimental analysis.

Data sets (Abbreviations)	Objects	Features	Classes
Wine recognition (WR)	178	13	3
Iris (Iris)	150	4	3
Spectf (SP)	267	44	2
Ecoli (Ec)	336	7	8
Gesture Phase A1 (GPA1)	1747	18	5
Pendigits (Pen)	10992	16	10

- USPS: The US Postal handwritten digits database contain 0 to 9 ten categories digits, each image is represented by 256-dimensional gray pixel values. The scale of it is 9298 Objects, 256 Dimensions and 10 Classes.
- YALE: The YALE data set contains 11 different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink, those are provided by 15 individuals. Each image is represented by 1024-dimensional gray pixel values. The scale of it is 165 Objects, 1024 Dimensions and 15 Classes.
- Caltech101-2 and Caltech101-7: Caltech101-2 and Caltech101-7 are the subsets of image dataset Caltech101 which contains 101 categories. Caltech101-2 contains two types of objects and each of them contains similar number of samples. Caltech101-7 contains 7 widely categories whose sample size varies greatly. Each image is represented by 1984-dimensional Histogram Oriented Gradient feature [52]. The scale of Caltech101-2 is 102 Objects, 1984 Dimensions and 2 Classes. The scale of Caltech101-7 is 1563 Objects, 1984 Dimensions and 7 Classes.

The Classes of every data set represent ground truth distribution of each data set which is used to verify whether the proposed method can reveal the inherent structure of data sets.

5.3.2. Clustering performance analysis

To evaluate the preponderance of proposed clustering algorithms, we first consider a widely used clustering performance index: accuracy (ACC), which can be regarded as a set matching method [31,53].

Accuracy: Clustering Accuracy criteria makes use of the true class labels to evaluate the clustering result distribution on each given data. It can be defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} I_{map(c_i)=l_i}}{n}$$

where c_i is the cluster label of x_i and l_i is the true class label, n is the total number of objects, $I_{x=y}$ denotes the indicator function that equals 1 if $x = y$ and equals 0 otherwise, $map(c_i)$ represents the permutation mapping function which best match the cluster label set and true label set.

The second widely used index is normalized mutual information (NMI), which evaluates the performance of clustering algorithms from a viewpoint of information gain.

NMI: Normalized Mutual Information is used to estimate the probability distribution of the clustering results [46]. Given the clustering result, the NMI can be computed by

$$NMI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} n_{i,j} log \frac{n_{i,j}}{n_i n_j}}{\sqrt{(\sum_{i=1}^{c} n_i log \frac{n_i}{n})(\sum_{j=1}^{c} n_j log \frac{n_j}{n})}}$$

where n_i is the number of data contained in the cluster $c_i (1 \le i \le c)$, n_j is the number of data belonging to the truth class $t_j (1 \le j \le c)$, and $n_{i,j}$ is the number of data in the intersection of c_i and t_j. The $NMI = 1$ when the clustering result distribution and real class are equivalent, otherwise, $NMI = 0$ when the compared clustering result distribution are absolutely different.

The Purity is another external criterion which attempts to measure the similarity between two clustering partitions of objects in the same data set.

Purity: Purity only considers the true positive data points that have been divided [54]. The special details as follows:

$$Purity = \sum_{i=1}^{c} \frac{max_j(n_i^j)}{n}$$

Table 4
M ± STD of ACC, NMI and Purity on UCI datasets.

Data sets	Methods	AC	NMI	Purity
WR	FCM	0.9551 ± 0.0000	0.8433 ± 0.0000	0.9551 ± 0.0000
	KFCM	0.9553 ± 0.0000	0.8511 ± 0.0000	0.9553 ± 0.0000
	PoFCM	0.9438 ± 0.0000	0.7969 ± 0.0000	0.9438 ± 0.0000
	DrFCM	0.9598 ± 0.0885	0.8577 ± 0.1319	0.9598 ± 0.0885
	DiFCM-lin	**0.9607 ± 0.0000**	0.8613 ± 0.0000	**0.9607 ± 0.0000**
	DiFCM-pol	**0.9607 ± 0.0000**	0.8681 ± 0.0000	**0.9607 ± 0.0000**
	DiFCM-sig	0.9507 ± 0.0000	**0.8781 ± 0.0000**	0.9507 ± 0.0000
Iris	FCM	0.9000 ± 0.0000	0.7540 ± 0.0000	0.9113 ± 0.0000
	KFCM	0.9160 ± 0.0084	0.7739 ± 0.0142	0.9160 ± 0.0084
	PoFCM	**0.9438 ± 0.0000**	**0.7969 ± 0.0000**	**0.9438 ± 0.0000**
	DrFCM	0.9270 ± 0.1335	0.7531 ± 0.2258	0.9127 ± 0.1335
	DiFCM-lin	0.9313 ± 0.0077	0.7940 ± 0.0000	0.9313 ± 0.0000
	DiFCM-pol	0.9000 ± 0.0000	0.7446 ± 0.0000	0.9000 ± 0.0000
	DiFCM-sig	0.8933 ± 0.0000	0.7371 ± 0.0000	0.8933 ± 0.0000
SP	FCM	0.7250 ± 0.0000	0.1900 ± 0.0001	0.7250 ± 0.0000
	KFCM	0.6875 ± 0.0000	0.2020 ± 0.0000	0.6875 ± 0.0000
	PoFCM	0.6963 ± 0.0664	0.3093 ± 0.0421	0.6963 ± 0.0664
	DrFCM	0.7510 ± 0.0444	**0.3350 ± 0.0656**	0.7475 ± 0.0444
	DiFCM-lin	**0.7587 ± 0.0318**	0.2384 ± 0.0581	**0.7587 ± 0.0318**
	DiFCM-pol	0.7525 ± 0.0348	0.2294 ± 0.0629	0.7525 ± 0.0348
	DiFCM-sig	0.7000 ± 0.0000	0.2365 ± 0.0000	0.7000 ± 0.0000
Ec	FCM	0.5161 ± 0.0068	0.4869 ± 0.0070	0.7970 ± 0.0238
	KFCM	0.6637 ± 0.0615	0.5021 ± 0.0523	0.7312 ± 0.0325
	PoFCM	0.6256 ± 0.0000	0.5253 ± 0.0010	0.7956 ± 0.0000
	DrFCM	0.6265 ± 0.1168	0.4339 ± 0.1667	0.6884 ± 0.1123
	DiFCM-lin	**0.7649 ± 0.0000**	**0.6091 ± 0.0007**	0.8217 ± 0.0151
	DiFCM-pol	0.6917 ± 0.0523	0.5354 ± 0.0197	**0.8289 ± 0.0040**
	DiFCM-sig	0.6454 ± 0.0077	0.4932 ± 0.0048	0.7818 ± 0.0028
GPA1	FCM	0.5026 ± 0.0000	0.2645 ± 0.0000	0.7207 ± 0.0000
	KFCM	0.6089 ± 0.0543	0.2693 ± 0.0208	0.7013 ± 0.0117
	PoFCM	0.7202 ± 0.0000	0.2954 ± 0.0000	0.7202 ± 0.0000
	DrFCM	0.7188 ± 0.0784	0.3045 ± 0.0640	**0.7682 ± 0.0895**
	DiFCM-lin	0.7145 ± 0.0000	0.3103 ± 0.0000	0.7298 ± 0.0000
	DiFCM-pol	0.7141 ± 0.0008	**0.3125 ± 0.0054**	0.7234 ± 0.0077
	DiFCM-sig	**0.7327 ± 0.0000**	0.3016 ± 0.0000	0.7327 ± 0.0000
Pen	FCM	0.7369 ± 0.0029	0.6645 ± 0.0007	0.7387 ± 0.0008
	KFCM	0.7203 ± 0.0479	0.6684 ± 0.0156	0.7417 ± 0.0279
	PoFCM	0.7480 ± 0.0447	0.6630 ± 0.0114	0.7463 ± 0.0291
	DrFCM	0.7402 ± 0.0343	**0.6959 ± 0.0374**	0.7376 ± 0.0372
	DiFCM-lin	**0.7521 ± 0.0056**	0.6765 ± 0.0030	**0.7521 ± 0.0050**
	DiFCM-pol	0.7425 ± 0.0332	0.6724 ± 0.0119	0.7517 ± 0.0217
	DiFCM-sig	0.7469 ± 0.0024	0.6660 ± 0.0005	0.7386 ± 0.0007

where c is the number of the clusters, and n is the total number of the data points, n_i^j is the number of the i-th input class that is assigned to the j-th cluster.

All of the above measures are in $[0, 1]$, the maximum value 1 represents the clustering result equals to its real partition. A higher value indicates a closer match between clustering partition and true class.

To ensure the impartial comparison based on three evaluation indices, it's necessary to provide a uniform environment condition. First, we set the number of clusters to be the same as the number of true classes of each given data set [47]. A clustering algorithm is well-behaved if it's clustering result closely matches the truth class distribution for a given data set. Second, we initialize the membership matrix by randomly generating under constraint (2) and carry out algorithms for 100 times on each data set to observe their performance from the statistical viewpoint. Third, we go through all parameters under each fuzzy index offered in Example for every data set and illustrate best average and standard deviation on three indices obtained by 20 repeated experiments.

The clustering results on three indexes over UCI data sets and image data sets have been reported in Table 4 and Table 5. For each data set, our algorithms with three different regularization items and the corresponding clustering results are below the dotted line. Best performances on three evaluation indexes of all compared algorithm have been enhanced with black. Compared with the traditional FCM, all other modified FCM methods perform better over most data sets except on Spectf and YALE data sets, but the DiFCM-lin performs better on these two special data sets, it indicates that the diversity information between different centers is benefit to improve the clustering performance, though the diversity information is introduced to the objective function of FCM by different means. Compared with PoFCM and DrFCM, the performances on all three evaluation indexes of our proposed methods is only inferior on Iris data set and the gap is also very narrow, it indicates the effectiveness of encouraging divers among clustering centers, meanwhile attests the validity of using independent statistics to measure the diversity information between clustering centers. As for our own algorithms with three types of regularization items, their performances vary from data to data. This is in line with objective laws, because the nonlinear

Table 5
M ± STD of ACC, NMI and Purity on image datasets.

Data sets	Methods	AC	NMI	Purity
JAFFE	FCM	0.9329 ± 0.0686	0.9587 ± 0.0254	0.9484 ± 0.0486
	KFCM	0.9681 ± 0.0529	0.9654 ± 0.0671	0.9690 ± 0.0528
	PoFCM	0.9480 ± 0.0556	0.9602 ± 0.0510	0.9575 ± 0.0507
	DrFCM	0.9670 ± 0.0320	0.9530 ± 0.0025	0.9400 ± 0.0210
	DiFCM-lin	**0.9765 ± 0.0000**	**0.9699 ± 0.0000**	**0.9765 ± 0.0000**
	DiFCM-pol	0.9615 ± 0.0519	**0.9699 ± 0.0173**	0.9685 ± 0.0366
	DiFCM-sig	**0.9765 ± 0.0000**	0.9614 ± 0.0000	**0.9765 ± 0.0000**
Handwritten	FCM	0.7611 ± 0.0686	0.7411 ± 0.0365	0.7821 ± 0.0536
	KFCM	0.7748 ± 0.0543	0.7190 ± 0.0404	0.7806 ± 0.0495
	PoFCM	0.7881 ± 0.0232	0.7497 ± 0.1158	0.7985 ± 0.1133
	DrFCM	**0.8394 ± 0.0359**	0.7445 ± 0.0349	0.7575 ± 0.0379
	DiFCM-lin	0.7955 ± 0.0447	**0.7670 ± 0.0238**	**0.8217 ± 0.0320**
	DiFCM-pol	0.7855 ± 0.0240	0.7470 ± 0.0208	0.8117 ± 0.0304
	DiFCM-sig	0.7760 ± 0.0270	0.7370 ± 0.0238	0.8017 ± 0.0320
USPS	FCM	0.6044 ± 0.0406	0.5627 ± 0.02083	0.6750 ± 0.0374
	KFCM	0.6688 ± 0.0000	0.6067 ± 0.0000	0.7324 ± 0.0000
	PoFCM	0.6802 ± 0.0295	**0.6800 ± 0.0501**	0.7140 ± 0.0461
	DrFCM	**0.7341 ± 0.0120**	0.6720 ± 0.0030	0.7012 ± 0.0220
	DiFCM-lin	0.6637 ± 0.0162	0.6075 ± 0.0076	0.7304 ± 0.0130
	DiFCM-pol	0.6657 ± 0.0034	0.6067 ± 0.0021	0.7316 ± 0.0033
	DiFCM-sig	0.6405 ± 0.0024	0.6324 ± 0.0032	**0.7437 ± 0.0015**
YALE	FCM	0.4885 ± 0.0208	0.5247 ± 0.0155	0.4952 ± 0.0167
	KFCM	0.4194 ± 0.0339	0.4895 ± 0.0269	0.4339 ± 0.0332
	PoFCM	0.3600 ± 0.0396	0.3690 ± 0.0501	0.3752 ± 0.0369
	DrFCM	0.3036 ± 0.0343	0.3618 ± 0.0294	0.3255 ± 0.0304
	DiFCM-lin	**0.5055 ± 0.0259**	**0.5363 ± 0.0135**	**0.5103 ± 0.0255**
	DiFCM-pol	0.4939 ± 0.0209	0.5322 ± 0.0124	0.5012 ± 0.0232
	DiFCM-sig	0.3885 ± 0.0028	0.3573 ± 0.0003	0.3939 ± 0.0284
Caltech101-2	FCM	0.6313 ± 0.0671	0.3804 ± 0.0730	0.6600 ± 0.0959
	KFCM	0.7226 ± 0.0678	0.4069 ± 0.0981	0.7226 ± 0.0678
	PoFCM	0.7652 ± 0.0000	0.3389 ± 0.0000	0.7652 ± 0.0000
	DrFCM	0.6870 ± 0.0980	0.4466 ± 0.0123	0.6878 ± 0.0963
	DiFCM-lin	0.7391 ± 0.0012	**0.4683 ± 0.0058**	0.7391 ± 0.0012
	DiFCM-pol	0.7217 ± 0.0000	0.3846 ± 0.0880	0.7217 ± 0.0000
	DiFCM-sig	**0.7826 ± 0.0000**	0.4604 ± 0.0000	**0.7826 ± 0.0000**
Caltech101-7	FCM	0.5859 ± 0.0249	0.3239 ± 0.0014	0.7075 ± 0.0090
	KFCM	0.5982 ± 0.0886	0.3250 ± 0.0681	0.6196 ± 0.0751
	PoFCM	0.6161 ± 0.0854	0.3139 ± 0.0550	0.6836 ± 0.0384
	DrFCM	0.6378 ± 0.0654	0.3332 ± 0.0721	0.6550 ± 0.0589
	DiFCM-lin	**0.6624 ± 0.0003**	0.2792 ± 0.0008	0.7103 ± 0.0003
	DiFCM-pol	0.6404 ± 0.0000	**0.3372 ± 0.0000**	0.7089 ± 0.0000
	DiFCM-sig	0.6340 ± 0.0000	0.3175 ± 0.0131	**0.7407 ± 0.0000**

modes of each data adaptation is really different. However, the performance of DiFCM-lin still precedes other two kernel cases on most data sets, we suggest that mining the diversity information between centers using independent statistics HSIC with linear kernel can be the first to use.

5.4. Convergence analysis

The convergence of proposed method has been proofed mathematically, in this section, we analyze convergence of DiFCM-lin in real datasets, the other two kernel cases have similar convergence. We carry out experiments on twelve data sets, the convergence curves on twelve data sets have been plotted in Fig. 5 and Fig. 6 by DiFCM-lin method with fuzzy index $m = 1.5$ and trade off parameter $\lambda = 1$. We can see the objective value converges for all data sets less than 100 iterations, it implies that the alternating optimization strategy to solve the proposed method is effective.

5.5. Sensitivity analysis

In our series of proposed algorithms, the number of parameters varies with the selected kernel function, which determines the computing form of diversity regularization. In this section, we still just display the impact of parameters on the performance of the DiFCM-lin algorithm. Two parameters are included in DiFCM-lin, the fuzzy index m and the trade off parameter λ, we fix one of them to analyze the other's influence to the clustering performance.

With fixed value $m = 1.1$, we plot sensitivity curves for trade off parameter λ in Fig. 7, where λ is assigned in $\{10^{-5}, \cdots, 10^{-1}, 10^{0}, 10^{1}, \cdots 10^{5}\}$. We can find that the clustering performances on ACC, NMI and Purity over all datasets have similar varying tendency with the increase of λ. Specifically, the ACC, NMI and Purity values slightly increase first and then decrease along with the increasing value of λ, however, the inflection point is different on different data sets, it

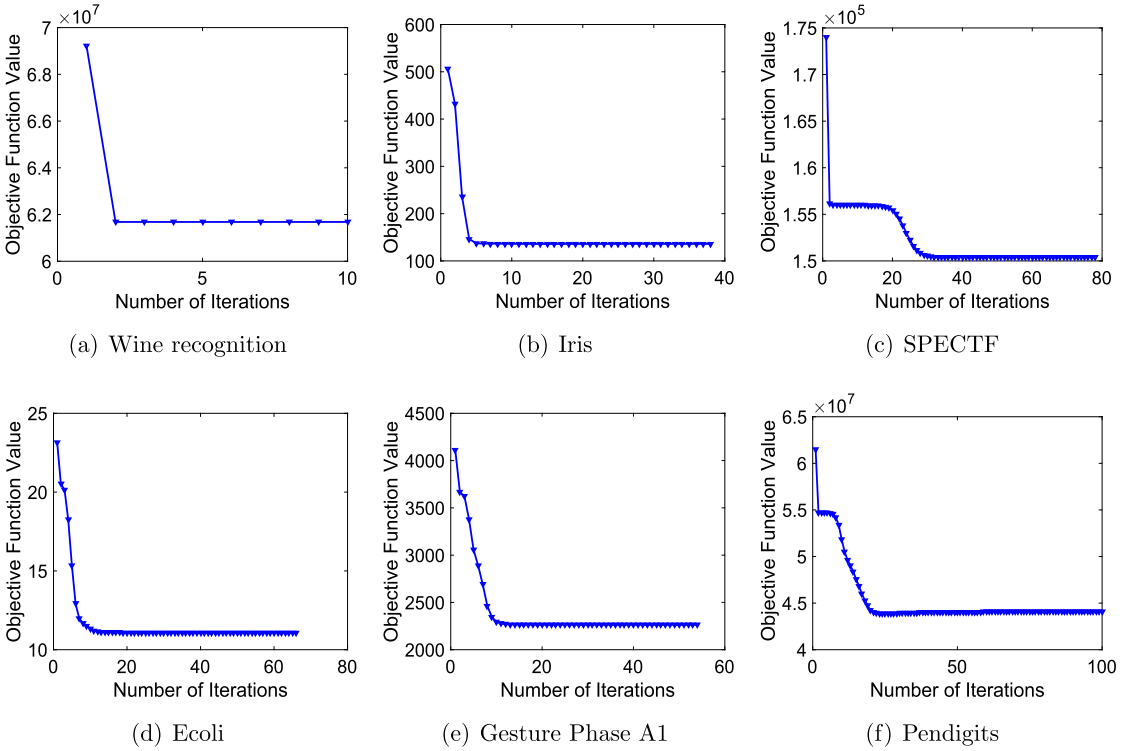


Fig. 5. Convergence performance of DiFCM-lin method on UCI datasets.

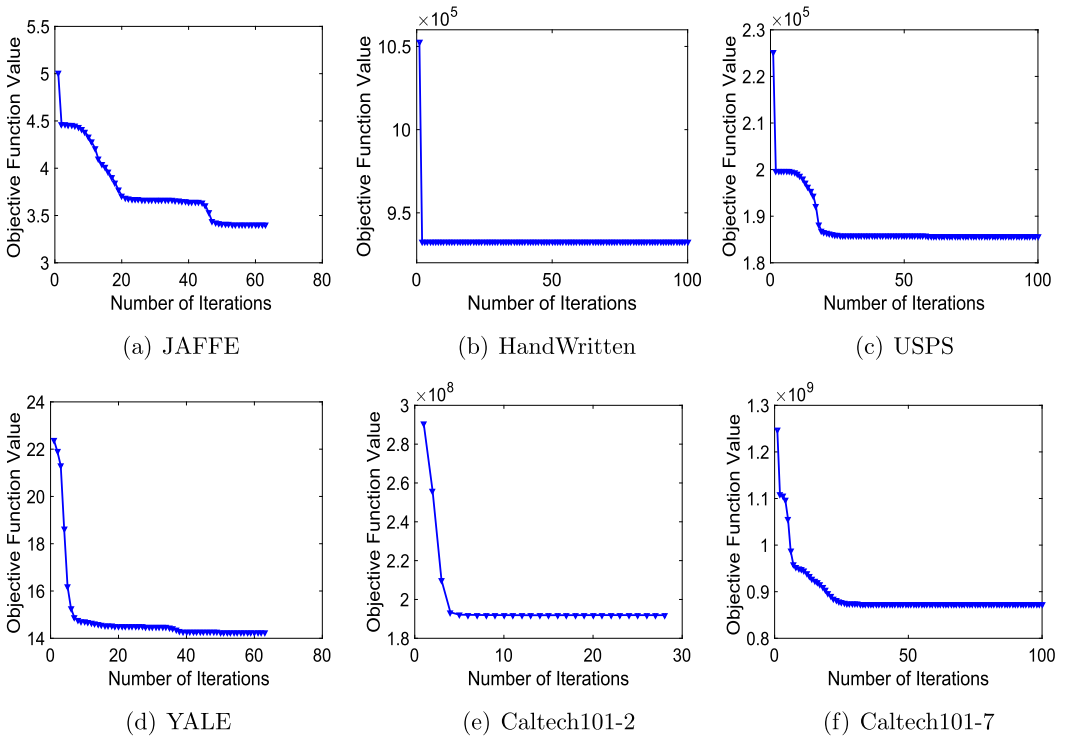
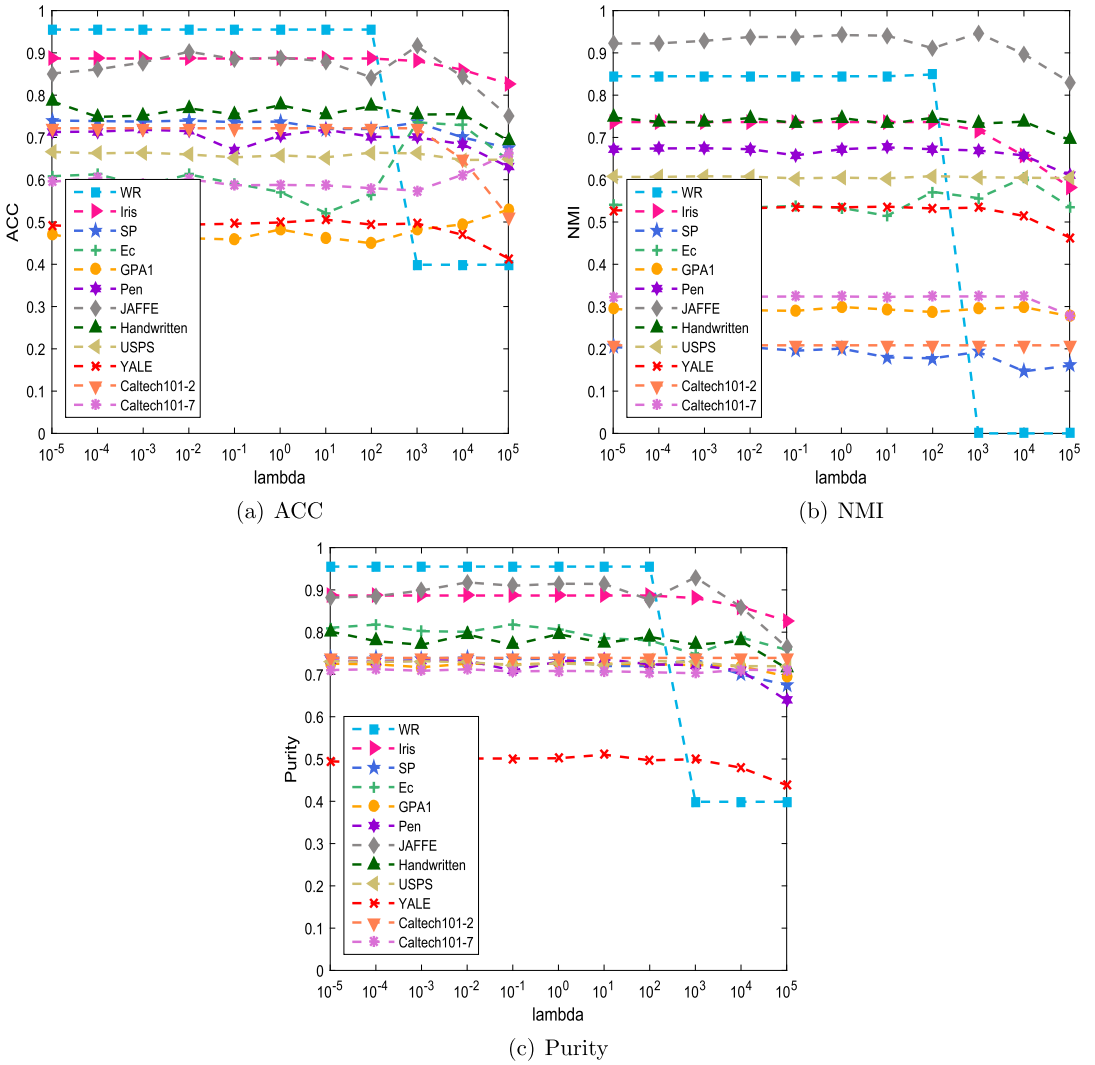


Fig. 6. Convergence performance of DiFCM-lin method on image datasets.



(a) ACC (b) NMI

(c) Purity

Fig. 7. Sensitivity analysis of λ in DiFCM-lin method.

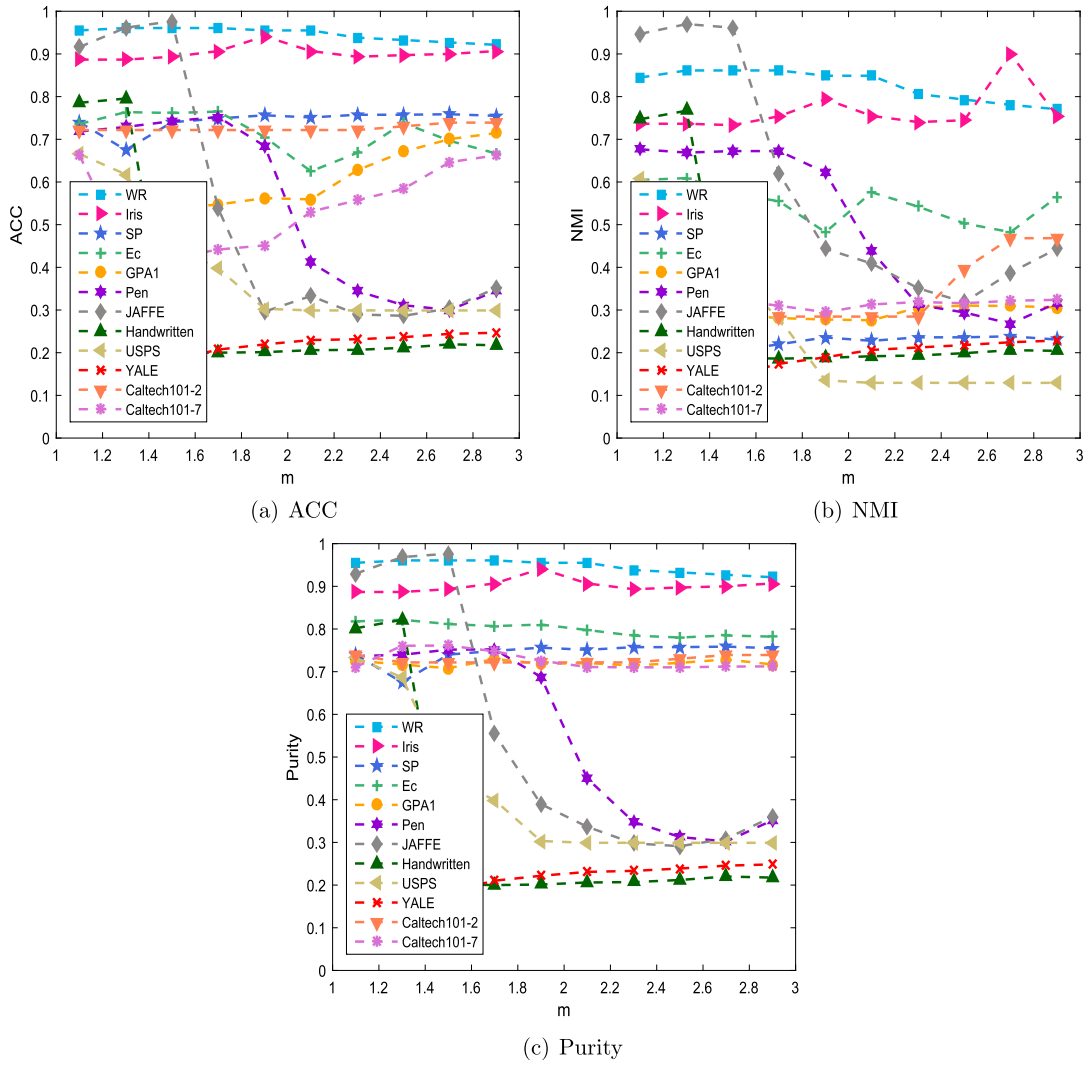
indicates that λ will effect the clustering performance, a proper value seriously depends on a special data set especially the performance on the Wine recognition data.

On the other hand, we evaluate the influence of fuzzy index m on the clustering results in Fig. 8, where m is tuned in $[1.1, 2.9]$ with step-size 0.2. We traverse all $\lambda \in \{10^{-5}, \cdots, 10^{-1}, 10^{0}, 10^{1}, \cdots 10^{5}\}$ and display the best performance on ACC, NMI and Purity over all data sets for per m. The present results show that fuzzy index has distinguishable effect on different data sets, but for the same data, the variation tendency on three index is nearing a consensus.

According to the above analysis, the clustering results of diversity-induced FCM are affected by the varying values of m and λ. But in general, the trade off parameter λ and fuzzy index m are respectively tuned in $[10^{-2}, 10^{2}]$ and $[1.1, 2]$, the clustering results remain relatively robust and approving.

6. Conclusions

This work induces a DiFCM framework based on fuzzy C-Means clustering, which takes the diversity information among clusters into consideration during the clustering process. Concretely, we combine the within-cluster similarity information and between-cluster diversity information simultaneously and address the objective function using alternating minimizing optimization strategy by decomposing it into two subproblems. With different kernel functions to measure the diversity information, the diversity regularized terms to the object function are different, so we have three specific forms of DiFCM and we provide the centers updating formulas respectively. The learned center attracts the points have the same group with it as well as excludes the impact from other clusters. It contributes to detect distinct clusters especially in the case due to existing overlap regions among clusters, it also helps to distinguish groups from unbalance datasets where the group



(a) ACC (b) NMI

(c) Purity

Fig. 8. Sensitivity analysis of m in DiFCM-lin method.

size of different clusters differs greatly. The convergence of our algorithm is proved theoretically and experimentally and the sensitivity of parameters in algorithm for clustering performance is also discussed. Finally, we carry out our three algorithms over six UCI data sets and six image datasets and compare against four baselines on three clustering evaluation indexes, experimental results have manifested the proposed algorithm is effective.

The HSIC measure has been proved to be effective to measure the diversity information, it will be interesting to apply other dependence measures to reveal more interesting nonlinear information. A good clustering algorithm is not only influenced by the size and shape of individual clusters, but also influenced by the relative position among different clusters. In this paper, we just attempt to model those factors based on the fuzzy C-Means clustering algorithm framework. In the future, we can restructure other famous clustering method or propose new generalization methods to improve the clustering performance.

Acknowledgements

We are very grateful to the anonymous reviewers for their valuable comments and suggestions, which significantly enhance the quality of this paper.

This work was supported by National Natural Science Fund of China (No. 61672332, 61322211, 61432011, U1435212, 61502289, 61872226), Program for New Century Excellent Talents in University (No. NCET-12-1031), Program for the Outstanding Innovative Teams of Higher Learning Institutions of Shanxi, Program for the Young San Jin Scholars of Shanxi, and Program for Natural Science Foundation of Shanxi Province (No. 201701D121052).

References

[1] Lotfi A. Zadeh, Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems, Soft Comput. 2 (1) (1998) 23–25.
[2] Jing Tao Yao, Athanasios V. Vasilakos, Witold Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.
[3] Pierpaolo D'Urso, Informational Paradigm, Management of Uncertainty and Theoretical Formalisms in the Clustering Framework, Elsevier Science Inc., 2017.
[4] Yuhua Qian, Hu Zhang, Feijiang Li, Qinghua Hu, Jiye Liang, Set-based granular computing: a lattice model, Int. J. Approx. Reason. 55 (3) (2014) 834–852.
[5] Yuhua Qian, Xinyan Liang, Guoping Lin, Qian Guo, Jiye Liang, Local multigranulation decision-theoretic rough sets, Int. J. Approx. Reason. 82 (2017) 119–137.
[6] Rui Xu, Donald Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.
[7] Anil K. Jain, Data clustering: 50 years beyond k-means, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2008, pp. 3–4.
[8] Ji Ye Liang, Yu Hua Qian, L.I. De Yu, H.U. Qing Hua, Theory and method of granular computing for big data mining, Sci. Sin. 45 (11) (2015) 1355.
[9] Richard O. Duda, Peter E. Hart, David G. Stork, et al., Pattern Classification, vol. 2, Wiley, New York, 1973.
[10] Scott D. Connell, Anil K. Jain, Writer adaptation for online handwriting recognition, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 329–346.
[11] Yuhua Qian, Honghong Cheng, Jieting Wang, Jiye Liang, Witold Pedrycz, Chuangyin Dang, Grouping granular structures in human granulation intelligence, Inf. Sci. 382 (2017) 150–169.
[12] Anil K. Jain, M. Narasimha Murty, Patrick J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.
[13] Cor J. Veenman, Marcel J.T. Reinders, Eric Backer, A maximum variance cluster algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 24 (9) (2002) 1273–1280.
[14] Wang Peizhuang, Pattern recognition with fuzzy objective function algorithms (James C. Bezdek), SIAM Rev. 25 (3) (1983) 442.
[15] Elizabeth Ann Maharaj, Pierpaolo D'Urso, Fuzzy clustering of time series in the frequency domain, Inf. Sci. 181 (7) (2011) 1187–1211.
[16] Maria Brigida Ferraro, Paolo Giordani, Possibilistic and fuzzy clustering methods for robust analysis of non-precise data, Int. J. Approx. Reason. 88 (2017).
[17] Alan Wee-Chung Liew, Hong Yan, An adaptive spatial fuzzy clustering algorithm for 3-d mr image segmentation, IEEE Trans. Med. Imaging 22 (9) (2003) 1063–1075.
[18] Doulaye Dembele, Philippe Kastner, Fuzzy c-means method for clustering microarray data, Bioinformatics 19 (8) (2003) 973–980.
[19] Pierpaolo D'Urso, Fuzzy clustering, in: Handbook of Cluster Analysis, CRC Press, 2015.
[20] Heiko Timm, Christian Borgelt, Christian Doring, Rudolf Kruse, An extension to possibilistic fuzzy cluster analysis, Fuzzy Sets Syst. 147 (1) (2004) 3–16.
[21] Yinghua Shen, Witold Pedrycz, Collaborative fuzzy clustering algorithm: some refinements, Int. J. Approx. Reason. 86 (2017).
[22] Weiling Cai, Songcan Chen, Daoqiang Zhang, Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation, Pattern Recognit. 40 (3) (2007) 825–838.
[23] Liang Bai, Jiye Liang, Chuangyin Dang, Fuyuan Cao, A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets Syst. 215 (2013) 55–73.
[24] Xuesong Yin, Ting Shu, Qi Huang, Semi-supervised fuzzy clustering with metric learning and entropy regularization, Knowl.-Based Syst. 35 (2012) 304–311.
[25] Liyong Zhang, Wanxie Zhong, Chongquan Zhong, Wei Lu, Xiaodong Liu, Witold Pedrycz, Fuzzy c-means clustering based on dual expression between cluster prototypes and reconstructed data, Int. J. Approx. Reason. 90 (2017).
[26] Lingling Zhang, Minnan Luo, Jun Liu, Zhihui Li, Qinghua Zheng, Diverse fuzzy c-means for image clustering, Pattern Recognit. Lett. (2018), https://doi.org/10.1016/j.patrec.2018.07.004.
[27] Peixin Hou, Jiguang Yue, Hao Deng, Shuguang Liu, An uncertainty perspective to pcm and apcm clustering, Int. J. Approx. Reason. 95 (2018).
[28] Arthur Gretton, Olivier Bousquet, Alex Smola, Bernhard Schölkopf, Measuring statistical dependence with Hilbert–Schmidt norms, in: International Conference on Algorithmic Learning Theory, Springer, 2005, pp. 63–77.
[29] Hyeong-Seog Kim, Joo-Hong Kim, Chang-Hoi Ho, Pao-Shin Chu, Pattern classification of typhoon tracks using the fuzzy c-means clustering method, J. Climate 24 (2) (2011) 488–508.
[30] Dimitri P. Bertsekas, Nonlinear Programming, Athena Scientific Belmont, 1999.
[31] Jin Huang, Feiping Nie, Heng Huang, Chris Ding, Robust manifold nonnegative matrix factorization, ACM Trans. Knowl. Discov. Data 8 (3) (2014) 11.
[32] Donglin Niu, Jennifer G. Dy, Michael I. Jordan, Multiple non-redundant spectral clustering views, in: Proceedings of the 27th International Conference on Machine Learning, ICML-10, 2010, pp. 831–838.
[33] Le Song, Alex Smola, Arthur Gretton, Karsten M. Borgwardt, A dependence maximization view of clustering, in: Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 815–822.
[34] Yale Chang, Junxiang Chen, Michael H. Cho, Peter J. Castaidi, Edwin K. Silverman, Jennifer G. Dy, Clustering with domain-specific usefulness scores, in: Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM, 2017, pp. 207–215.
[35] Beatriz Bueno Larraz, et al., Independence Measures, Master's thesis, 2015.
[36] Yang Zhao, Yong Dou, Xinwang Liu, Teng Li, A novel multi-view clustering method via low-rank and matrix-induced regularization, Neurocomputing 216 (2016) 342–350.
[37] James R. Schott, Matrix Analysis for Statistics, John Wiley & Sons, 2016.
[38] Pasi Fränti, Olli Virmajoki, Iterative shrinking method for clustering problems, Pattern Recognit. 39 (5) (2006) 761–775.
[39] Frank Hoppner, Frank Klawonn, A contribution to convergence theory of fuzzy c-means and derivatives, IEEE Trans. Fuzzy Syst. 11 (5) (2003) 682–694.
[40] L. Groll, J. Jakel, A new convergence proof of fuzzy c-means, IEEE Trans. Fuzzy Syst. 13 (5) (2005) 717–720.
[41] James C. Bezdek, Richard J. Hathaway, Michael J. Sabin, William T. Tucker, Convergence theory for fuzzy c-means: counterexamples and repairs, IEEE Trans. Syst. Man Cybern. 17 (5) (1987) 873–877.
[42] Richard J. Hathaway, James C. Bezdek, Local convergence of the fuzzy c-means algorithms, Pattern Recognit. 19 (6) (1986) 477–480.
[43] Yi Ding, Xian Fu, Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm, Neurocomputing 188 (2016) 233–238.
[44] Kuo-Lung Wu, Analysis of parameter selections for fuzzy c-means, Pattern Recognit. 45 (1) (2012) 407–415.
[45] Howon Choe, Jay B. Jordan, On the optimal choice of parameters in a fuzzy c-means algorithm, in: Fuzzy Systems, 1992, IEEE International Conference on, IEEE, 1992, pp. 349–354.
[46] Ibrahim Ozkan, I.B. Turksen, Upper and lower values for the level of fuzziness in fcm, Inf. Sci. 177 (23) (2007) 5143–5152.
[47] Ming Huang, Zhixun Xia, Hongbo Wang, Qinghua Zeng, Qian Wang, The range of the value for the fuzzifier of the fuzzy c-means algorithm, Pattern Recognit. Lett. 33 (16) (2012) 2280–2284.
[48] James Breneman, Kernel methods for pattern analysis, John Shawe-Taylor and Nello Cristianini, J. Am. Stat. Assoc. 101 (December) (2006) 1730.
[49] Mauricio A. Ivarez, Lorenzo Rosasco, Neil D. Lawrence, Kernels for vector-valued functions: a review, Found. Trends Mach. Learn. 4 (3) (2011) 195–266.
[50] Arthur Asuncion, David Newman, Uci Machine Learning Repository, 2007.

[51] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, Hong Jiang Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[52] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[53] Yuhua Qian, Feijiang Li, Jiye Liang, Bing Liu, Chuangyin Dang, Space structure and clustering of categorical data, IEEE Trans. Neural Netw. Learn. Syst. 27 (10) (2016) 2047–2059.
[54] Babak Rezaee, A cluster validity index for fuzzy clustering, Fuzzy Sets Syst. 161 (23) (2010) 3014–3025.